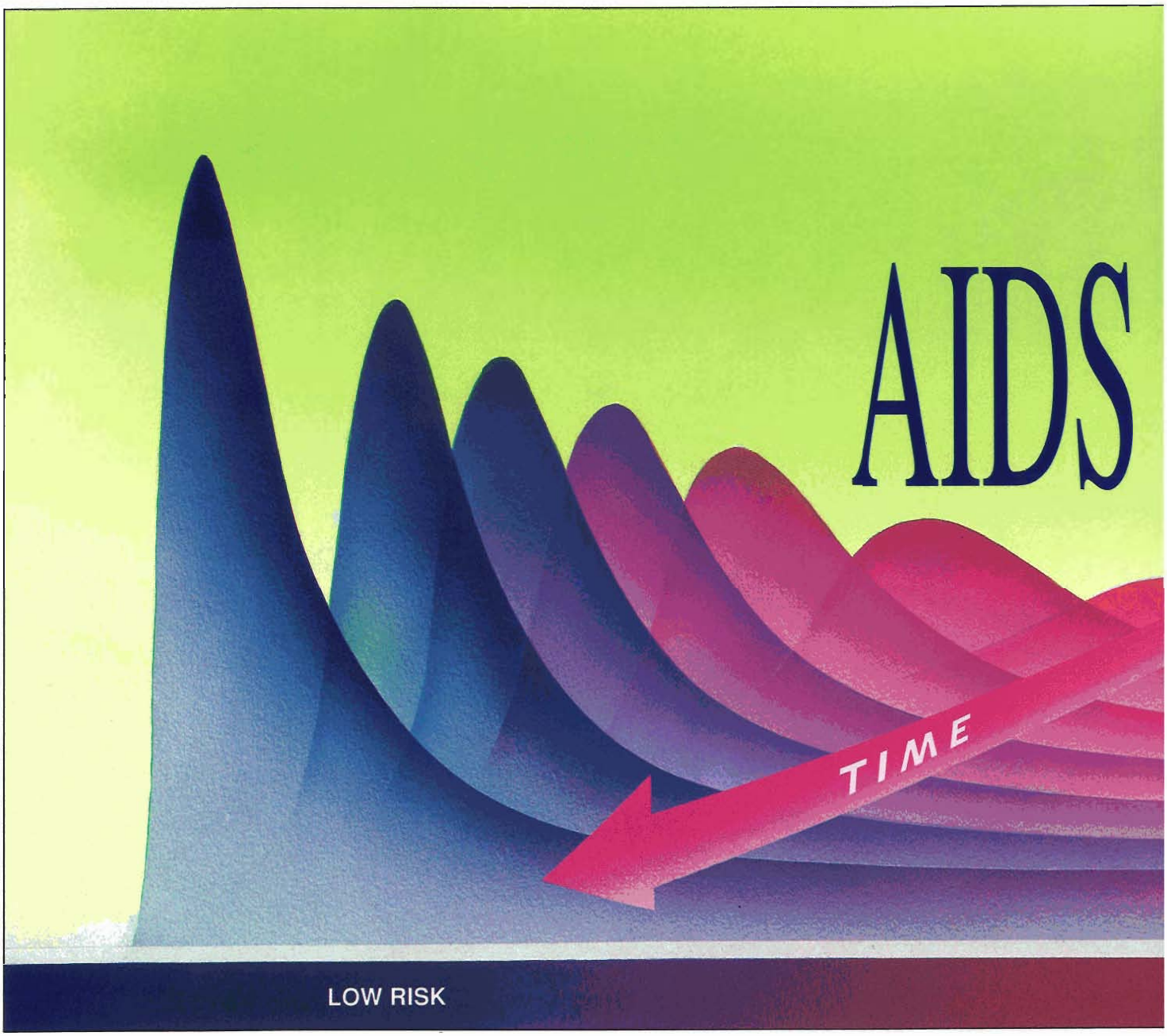


The biased-mixing risk-based model predicts that HIV infection spreads through the population from high-risk (red) to low-risk (blue) groups. The graph shows number of people infected (vertical scale) versus risk (left to right) versus time (back to front). As time increases, the peaks occur at lower and lower values of risk behavior. Also, they increase in height because there are many more people in low-risk than in high-risk groups.

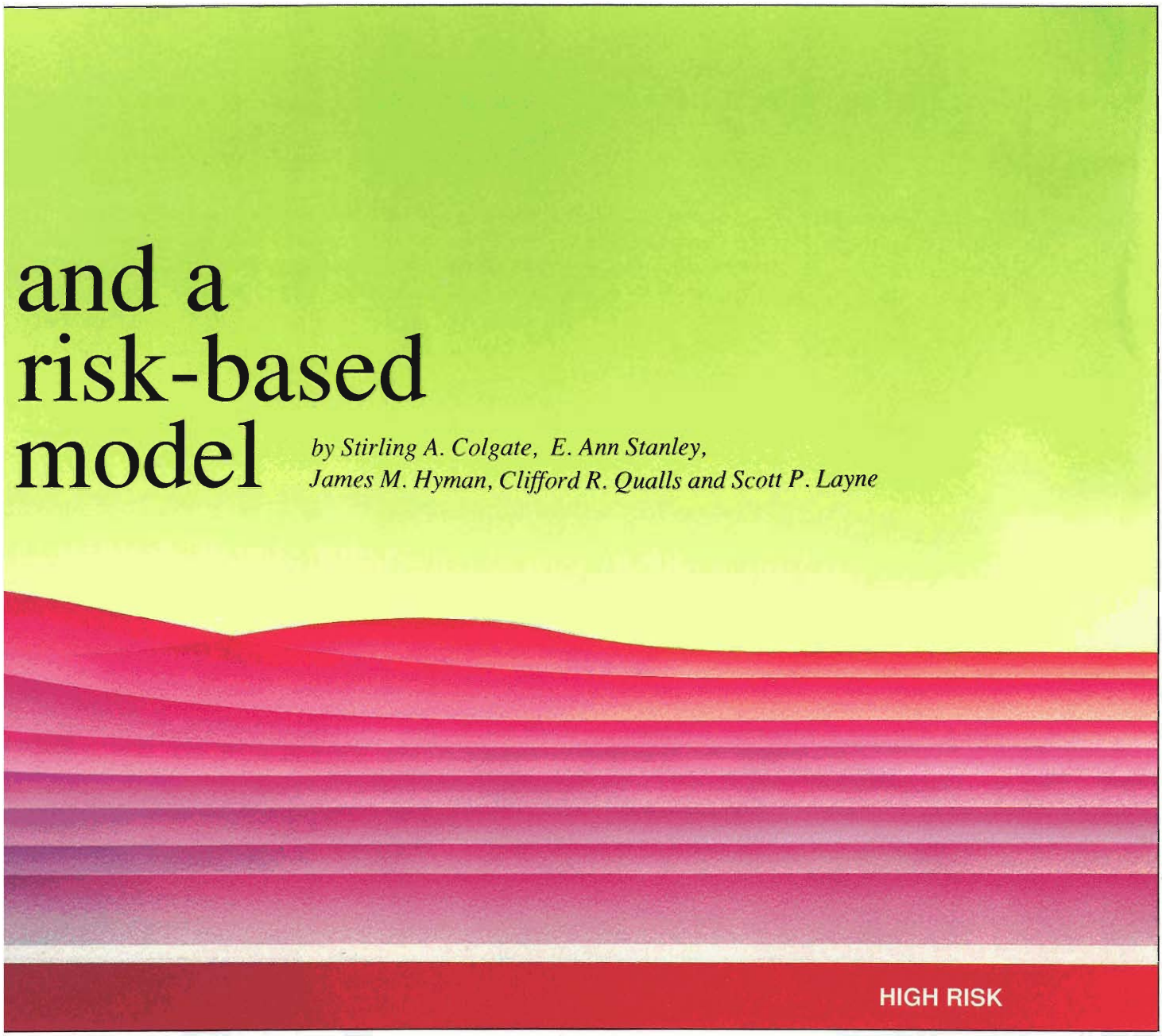
The threat of AIDS looms ominously over society. It has already devastated the male-homosexual population and is spreading rapidly among intravenous-drug users, through sexual contact to their partners and through perinatal contact to their children. Although many promising therapies are on the horizon, it appears unlikely that a definitive cure or preventive vaccine will ever be developed. Also, we don't know where this lethal disease will spread next and whether it will reach epidemic proportions among the bulk of our population.

In an effort to make a quantitative assessment of the threat, we adopted the philosophy that to predict the future we need to understand the past. How has the number of AIDS cases grown over time? How has the number grown among various subgroups of the population? What risk behavior is correlated with becoming infected? How does the long and variable time between infection and appearance of symptoms affect the spread of the disease? Can the known data be used to make a plausible model that agrees with the history of the epidemic to date?

We began our effort by looking at the most reliable data on the course of the epidemic—those compiled by the Centers for Disease Control (CDC) on the total number of AIDS cases in the United States as a function of time.

and a risk-based model

by Stirling A. Colgate, E. Ann Stanley, James M. Hyman, Clifford R. Qualls and Scott P. Layne



Airbrush art by David Delano

Because the United States has a large number of AIDS cases and a legal requirement for reporting them, the CDC data are the most statistically significant available.

Analysis of the United States data revealed two surprising facts. First, the number of cases has not grown exponentially with time, but rather cubically with time, or as t^3. The difference may not appear significant until one realizes that previous epidemiological models for the spread of diseases predict exponential growth during the early phases of an epidemic, and further, most epidemics so far studied have followed that pattern. The second surprise came when the data were broken down into subgroups by race and sex or sexual preference. Again the number of cases in each subgroup grew as t^3, and further, the cubic growth for each group appeared to start at nearly the same time.

The model we present here was developed to explain the cubic growth of AIDS cases in the United States. It builds on the fact that the level of "risky" behavior—in particular the sexual behavior that puts one at risk of contracting the AIDS virus—varies among the population according to a distribution that we speculate may be universal for all populations. Thus it is called a risk-based model. It also depends on another assumption about human behavior, namely, that people

with similar risk behavior tend to mix, or interact, primarily among themselves (biased mixing) rather than randomly with everyone (homogeneous mixing). The details of our risk-based, biased-mixing model form a logical, coherent framework for interpreting the currently available data for the United States, but before we launch into details we want to emphasize one critical insight.

Since the growth in number of AIDS cases is cubic, the doubling time for the epidemic (the time for the number of cases to double) is continuously increasing. By contrast, if the growth were exponential, the doubling time would remain constant. In the framework of standard epidemiological models, the observed lengthening of the doubling time for AIDS (and hence its decreasing relative growth rate) might be attributed to changes in people's sexual behavior as a result of learning about AIDS. That interpretation has been promulgated in the press and has fostered complacency about the efficacy of education. Unfortunately, it is false because the long incubation time from infection to AIDS means that the effects of learning could not have been seen in the data until very recently.

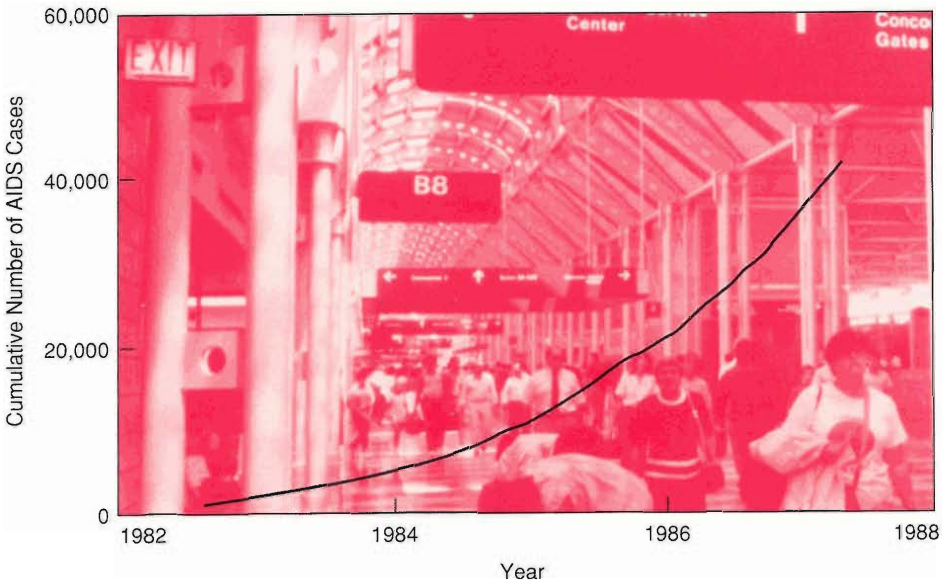
The people who developed AIDS in the early to mid 1980s were infected with the virus that causes AIDS (the human immunodeficiency virus, or HIV) in the late 1970s and early 1980s, long before learning could have affected a major fraction of the male-homosexual population. So behavior changes, if any, could not have been nearly enough to give cubic growth of AIDS in the late 1970s and early 1980s. Thus the impact of learning cannot explain the observed cubic growth. Another possibility to consider is that the combined effect (or convolution) of an exponential growth in HIV infections and a highly variable time for conversion from infection to AIDS yields a power law. After an initial transient, however, an exponential convoluted with any bounded conversion function is still an exponential, not a power law. Moreover, it is unlikely that the initial transient would have the long, clearly defined cubic behavior seen in the data.

We have looked with considerable diligence for possible causes of cubic growth other than behavioral changes due to learning. We have concluded that the risk-based, biased-mixing model presented here best fits the observations. Our model is an extension of an earlier risk-based model of May and Anderson. They assumed homogeneous rather than biased mixing of the susceptible population and so predicted an exponential for the early stages of the epidemic. We have drawn much from their work, but it was the contradiction between the theoretically nearly inevitable early exponential growth and the observed cubic growth that led us to the following biased-mixing model. We also realized that random mixing is sociologically unrealistic.

The general mathematical formalism for our model is presented in "Mathematical Formalism for the Risk-Based Model of AIDS." Numerical solutions for different assumptions about population mixing and variability of infectiousness are presented in "Numerical Results of the Risk-Based Model." Here we will present an intuitive and simplified version of the model that emphasizes the main features leading to cubic growth, the quantitative predictions of the model, and the questions about human behavior and HIV transmission that must be answered before we can determine whether the patterns we have identified for the past will continue in the future.

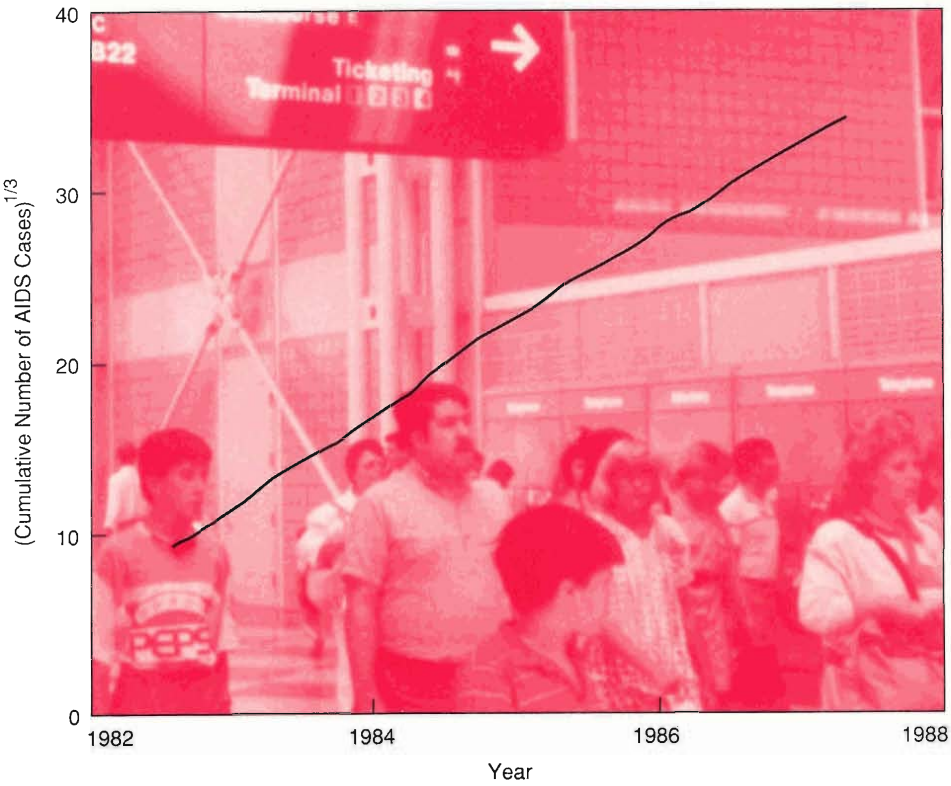
Cubic Growth of AIDS

The CDC data on cumulative number of AIDS cases in the United States between mid 1982 and early 1987 are shown in Fig. 1. Data for times prior to 1982.5 are not shown because they are statistically unreliable. Data collected since 1987.25 are also not shown because the surveillance definition of AIDS was changed in 1987. The effects of that change on reporting delays and/or on the cumulative number of AIDS cases have not been fully determined, but preliminary analysis suggests that



GROWTH OF AIDS IN THE U.S.

Fig. 1. Cumulative number of AIDS cases reported to the Centers for Disease Control through mid-1987. Data for times before mid-1982 are statistically unreliable and therefore not shown. More recent data have yet to be adjusted for the CDC's change in the definition of AIDS in May 1987.



CUBIC GROWTH OF AIDS

Fig. 2. The near linearity of this cube-root plot of the data shown in Fig. 1 indicates that the cumulative number of AIDS cases can be well represented by a cubic polynomial. We found that the best cubic fit is $A(t) = 174.6(t - 1981.2)^3 + 340$, where $A(t)$ is the cumulative number of AIDS cases and t is the yearly date. That fit reproduces the data to within 2 per cent.

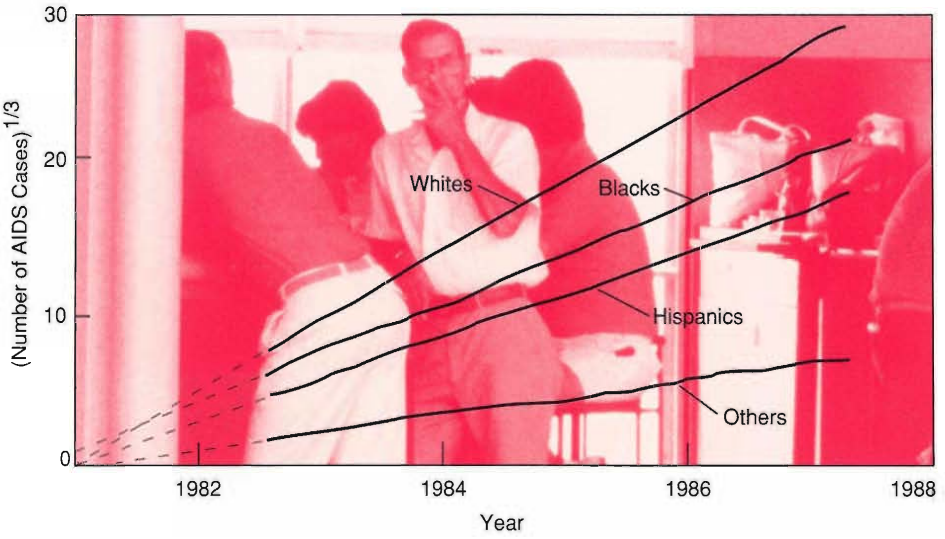
AIDS cases have continued to increase in a regular manner. The best fit to the data in Fig. 1 is the cubic function

$$A = 174.6(t_y - 1981.2)^3 + 340, \quad (1)$$

where A is the cumulative number of AIDS cases and t_y is the date in years. All the constants, including the date 1981.2, were determined by the statistical analysis. The cubic fit reproduces the data between 1982.5 and 1987.25 to within an accuracy of 2 per cent.

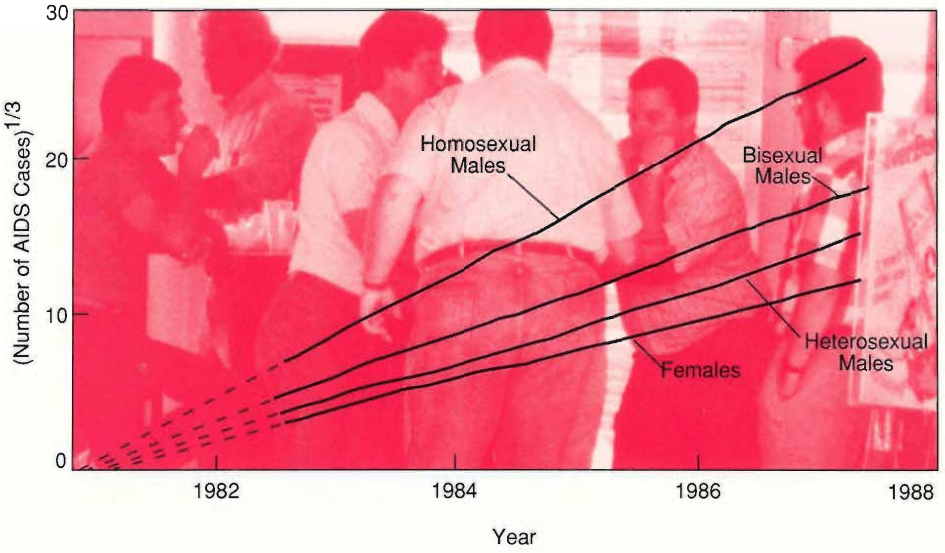
BREAKDOWN OF AIDS CASES BY RACE

Fig. 3. Plots of the cube root of the cumulative number of AIDS cases among homosexual, bisexual, and heterosexual males and among females are all nearly straight lines, indicating that AIDS has grown cubically in each group. Further, extrapolation of the curves (dashed lines) indicates that cubic growth began in all groups at approximately the same time that it began in the population as a whole. (Intravenous drug users have not been removed from the subgroups and the heterosexual-male subgroup includes cases of transfusion-related AIDS among juvenile males.)



BREAKDOWN OF AIDS CASES BY SEX AND SEXUAL PREFERENCE

Fig. 4. Plots of the cube root of the cumulative number of AIDS cases among whites, blacks, Hispanics, and other racial groups again indicate cubic growth of AIDS beginning about 1981. Breakdowns of the data by geographic categories also show cubic growth.



To show the cubic growth more clearly we plot the cube root of the total number of AIDS cases in Fig. 2 and see surprisingly small deviations from a straight line, of the order a few per cent for $t_y \geq 1982.5$. Hence, if we measure time t from 1981.2 and neglect the small error, then

$$A = A_0 + A_1 t^3, \quad \text{for} \quad t > 1.3 \ \text{years}, \tag{2}$$

where $A_0 = 340$ and $A_1 = 174.6$. Thus, after an initial transient, AIDS cases grow as the cube of time.

Breakdowns of the data by sex or sexual preference (Fig. 3) and race (Fig. 4) again show the same form of cubic growth for each subgroup. This is surprising, since the fact that the sum of all the data is cubic requires that the separate cubics be synchronized in time, in this case to within less than six months. In addition to presenting our model for cubic growth, we will discuss a possible seeding process for the initial cases of AIDS (see "The Seeding Wave") that is consistent with both the assumptions of our model and the synchronization of cubic growth in various subgroups.

Currently (third quarter of 1988), the CDC reported a cumulative total of 74,904 AIDS cases under their expanded mid-1987 surveillance definition. Of those about 14 per cent fell under the new categories added in mid-1987. More difficult to determine are the effects of delays between diagnosis and reporting to the CDC caused by the redefinition. The median reporting delay prior to the redefinition was about 3 months, and adjustments made for those delays have visible effects 36 months into the past on a graph such as the graph shown in Fig. 1. After the redefinition in mid-1987, the median reporting delay lengthened to about two years, and the reporting situation is still in transition. Consequently, we must await further data before we can model the effects of the transient caused by the redefinition and determine whether or not cubic growth has continued to the present. Nevertheless, we can say with certainty that the growth in AIDS cases is still polynomial of degree less than 4.

Expected Exponential Growth

We start by showing that the initial growth of AIDS (or of any infectious disease) would be exponential provided the population was homogeneous and did not change its behavior. We assume AIDS is the long-term result of infection by HIV and derive an equation for the rate of growth in the number of infected persons. Let I be the number of persons infected at time t in a population of size N. Assume that α, the rate at which an infected person transmits the AIDS virus to others, does not vary with time nor from person to person. Then during the time interval dt the I infected persons in the population would infect αI persons. But the fraction I/N of those αI persons are already infected, and so the number of additional persons infected during dt is $\alpha I - \alpha I(I/N)$; that is,

$$\frac{dI}{dt} = \alpha I \left(1 - \frac{I}{N}\right). \tag{3}$$

Equation 3 is called a logistic equation (or an equation of mass action) and is the basic equation of epidemiology. During the initial phases of the epidemic, $1 - \frac{I}{N}$ is approximately 1, and we can approximate Eq. 3 by

$$dI/dt = \alpha I. \tag{4}$$

Equation 4 has the exponential solution

$$I = I_1 e^{\alpha t}, \tag{5}$$

where I_1 is the number infected at $t = 0$.

Exponential growth is characteristic of the initial phase of many epidemics and is a solution of many current AIDS models. Note that exponential growth implies a constant relative growth rate:

$$\text{relative growth rate} \equiv \frac{dI/dt}{I} = \alpha. \tag{6}$$

The number infected will continue to grow exponentially until the fraction infected is no longer small compared to unity. The population is then said to be approaching saturation, and the relative growth rate decreases. However, we have observed that even the first few thousand AIDS cases show cubic, not exponential, growth, so saturation of the population cannot be the explanation for the decrease in the relative growth rate.

What Makes a Power Law? Suppose now that the relative growth rate α is not constant in time but instead decreases inversely with time:

$$\alpha = \frac{m}{t}, \tag{7}$$

where m is a constant. Then Eq. 4 becomes

$$\frac{dI}{dt} = m\frac{I}{t}. \tag{8}$$

Equation 8 has a power-law solution, namely,

$$I = I_1 t^m; \tag{9}$$

that is, the number infected grows as the mth power of time. Moreover, since the doubling time t_d is inversely proportional to the relative growth rate m/t, t_d increases proportionally to t. In particular, $t_d = (\sqrt[m]{2} - 1)t$. The growth of AIDS is cubic, so $m = 3$ and $t_d = (\sqrt[3]{2} - 1)t \approx 0.26t$. The observed doubling time for the AIDS epidemic has increased linearly from less than 0.5 year to the current value of more than 2 years. That change in doubling time and relative growth rate (by more than a factor of 4) is dramatically different from the constant doubling time characteristic of exponential growth.

A Risk-Based Model

Any model for the spread of an infectious disease must take into account the mechanism of its transmission, the pattern of mixing among the population, and the infectiousness, or probability of transmission per contact. The primary mechanisms for transmitting the AIDS virus are sexual contact and sharing of intravenous needles among drug users. Since little is known about needle-sharing habits, we concentrate on transmission through sexual contact. Here we build on data from the homosexual and heterosexual community. The relative growth rate of infection α can be approximated as the product of three factors: the infectiousness i, or probability of infection per sexual contact with an infected person; the average number of sexual contacts per partner c; and the average number of new partners per time interval p. That is,

$$\alpha \approx icp. \tag{10}$$

Each of the factors in Eq. 10 can be a complicated function. For example, data suggest that infectiousness i is, on average, between 0.01 and 0.001 and that it varies with time since infection and, perhaps, from individual to individual (more about that later). The new-partner rate and the average number of contacts per partner certainly vary among the population and may depend on age, place of residence, race, personal history, and more. The general model presented in "Mathematical Formalism for the Risk-Based Model of AIDS" allows for some of these variations, but here we pick out the simplest features that lead to cubic growth.

The first crucial assumption of the risk-based model is that the susceptible population is divided into groups according to level of engaging in behavior that can lead to infection. The risk behavior most often correlated with HIV infection in the male-homosexual population (as suggested by the early work of the CDC) is frequent change of sexual partner, which we quantify as new-partner rate. The other behavior we consider is frequency of sexual contact (which is equal to the product cp in Eq. 10). Both sexual contact and some new partners are necessary to cause the epidemic.

If our model is to agree with observation, we must assume that the members of each risk group (whether the risk be new-partner rate or sexual-contact frequency) interact primarily, but not exclusively, among themselves; in other words, the mixing among the population as a whole is biased. We also assume that mixing within each risk group is homogeneous and that the relative growth rate α is proportional to the risk behavior r,

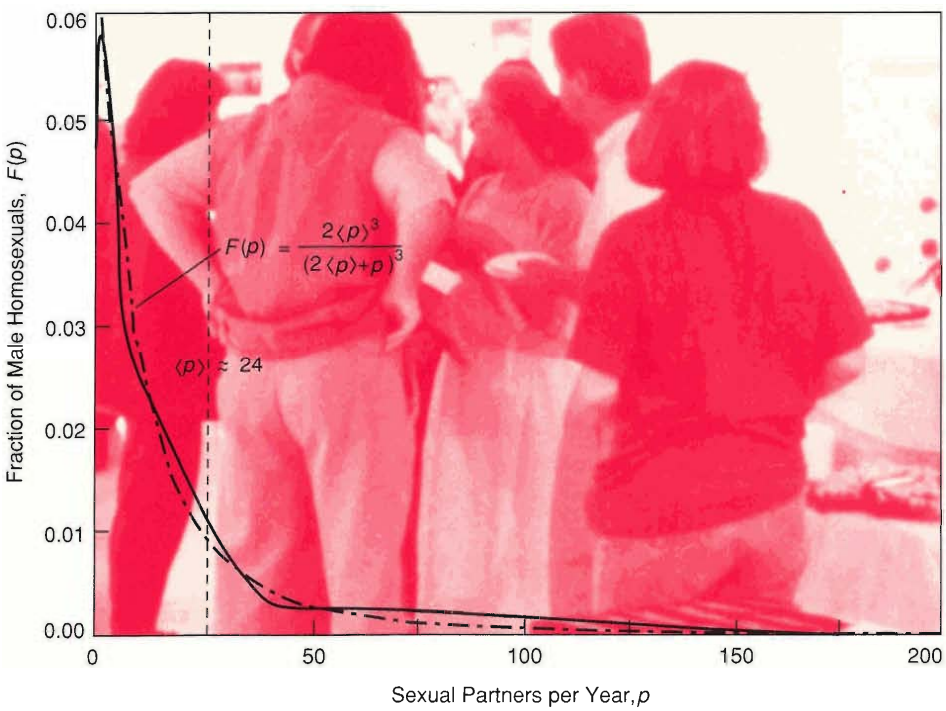
$$\alpha = \alpha' r, \tag{11}$$

so that infectiousness i is approximated as a constant.

Finally we assume (and justify below) that risk behavior is distributed among the high-risk groups as r^{-3}; that is, the number of people with risk behavior r, $N(r)$, is given by $N(r) \propto r^{-3}$. We believe that these assumptions are sufficient to explain the cubic growth of the AIDS epidemic. For the purposes of the model, it makes no difference what the risk behavior actually is—only that such a behavior exists and is distributed approximately as r^{-3}. However, because of past preconceptions and universal interest, we discuss the available data on the distribution of both new-partner rate and sexual-contact frequency. In doing so, we restrict ourselves primarily to cases of AIDS among homosexuals, which constitute roughly 65 per cent of the total number of cases. Our model can be applied to intravenous-drug users only when additional risk-behavior data are available.

Distribution of Risk Behavior

New-Partner Rate among Male Homosexuals. The best available data on new partner rate p come from studies of homosexual men. Although those data are usually presented in summary form (number with 20 to 40 partners in the past year, for example) and the sizes of the study samples tend to be small, all of the studies find similar distributions. The standard deviation σ is always larger than the mean $\langle p \rangle$, sometimes much larger. In other words, the population is not clustered about the mean but rather varies widely in its behavior. Moreover, a good fit to the data for



DISTRIBUTION OF NEW-PARTNER RATE

Fig. 5. A plot of $F(p)$, the fraction of a group of male homosexuals that had p sexual partners per year, versus p. Members of the group were attendees at London clinics for sexually transmitted diseases. (For more details about the data, see May and Anderson.) Also shown is our inverse-cubic fit to the data, $F(p) = 2\langle p \rangle^3/(p + \langle p \rangle)^3$, where $\langle p \rangle$ is the mean number of partners for the whole group.

p greater than a few partners per year is the distribution $p^{-\beta}$, where β is between 3 and 4. Figure 5 shows combined data from two studies of homosexual men attending London clinics for sexually transmitted diseases. Also shown is our cubic fit to the data $2\langle p\rangle^3/(\langle p\rangle + p)^3$. The two London studies are biased away from low-activity homosexual men; more randomly chosen samples tend to exhibit a $p^{-\beta}$ distribution at large p, but larger fractions of the samples lie at low p.

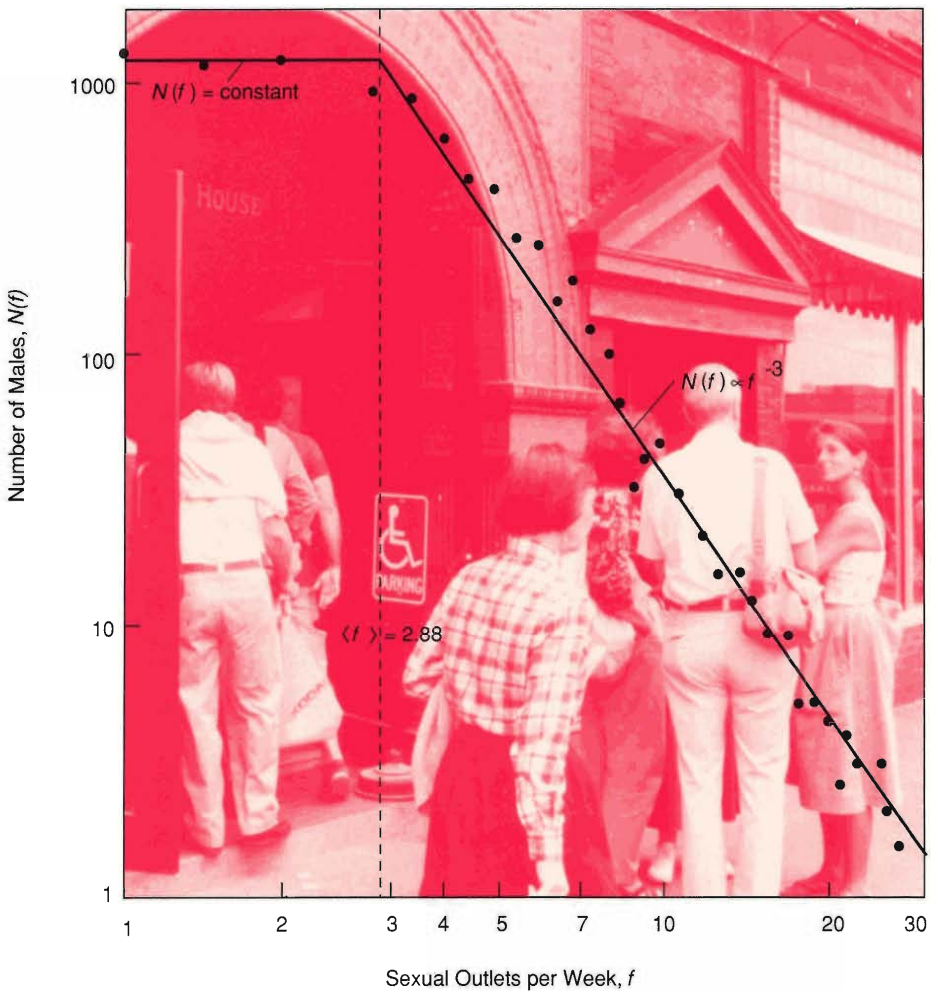
Either the published data are too crude (especially since the maximum value for the highest bin tends to be omitted) or the sample sizes are too small to distinguish between $\beta = 3$ and $\beta = 4$. In this paper we have chosen to use $\beta = 3$ to be consistent with the sexual-outlet frequency data of Kinsey, Pomeroy, and Martin (see below). The choice is important because in our model the value of β determines the growth rate of the epidemic (AIDS cases increase as t^{β}).

(One way to test the hypothesis that $\beta = 3$ for male homosexuals is to determine how the standard deviation σ of the distribution varies with sample size. If p is distributed as p^{-3}, then $\sigma/\langle p\rangle$ will increase as the sample size increases. By contrast, if p is distributed as p^{-4}, then $\sigma/\langle p\rangle$ will approach a limiting value of 4 as the sample size increases. Unfortunately, the data available are insufficient for us to apply this test.)

Sexual-Contact Frequency among Males. We now turn to the distribution of sexual-contact frequency. For that information we must rely on the data published in 1948 by Kinsey, Pomeroy, and Martin on sexual-outlet frequency among 11,467

DISTRIBUTION OF SEXUAL OUTLET FREQUENCY

Fig. 6. A plot of $N(f)$, the number of males among a large study group that had f sexual outlets per week, versus f. Also shown (solid line) is a distribution that fits the data well: $N(f)$ = constant for $f < \langle f\rangle$ (the mean sexual-outlet frequency) and $N(f) \propto f^{-3}$ for $f \geq \langle f\rangle$. The data are those of Kinsey, Pomeroy, and Martin for a group of 11,467 American males ranging in age from adolescence to thirty years.



American males ranging in age from adolescence to thirty years (Fig. 6). (The sexual outlets considered by Kinsey et al. include activities, such as masturbation, that are of little relevance to the spread of HIV infection. However, data more appropriate to our needs are not available.) We found that the Kinsey data could be well fit with a distribution similar to the distribution of new-partner rate among homosexual men. For values of sexual-outlet frequency f above the mean, the number of males at each f value, $N(f)$, is proportional to f^{-3}. The entire distribution is given by

$$\frac{N}{N_0} = \begin{cases} 1 & \text{if } f < \langle f \rangle \\ \left(\frac{f}{\langle f \rangle}\right)^{-3} & \text{if } f \geq \langle f \rangle, \end{cases} \tag{12}$$

where $\frac{3}{2}N_0$ is the sample size and $\langle f \rangle$ is the mean value of f.

The Kinsey data showed that sexual preference is independent of sexual-outlet frequency. That fact supports applying inverse cubic distributions to distinct sexual-preference groups, for example, to male homosexuals.

One may speculate that an inverse-cubic distribution of sexual-outlet frequency, $N \propto f^{-3}$, is a Darwinian barrier in behavior space produced by competition for a finite resource. If so, the distribution is not determined by a particular set of environmental or social influences but rather may be hard-wired into our genetic make-up. In any case, we find that both the distribution of sexual-outlet frequency among American males and the distribution of new-partner rate among a limited population of British homosexuals are described by inverse cubics. (That result suggests that an inverse cubic distribution of risk may also describe the heterosexual population.)

Sexual-Contact Frequency versus New-Partner Rate—Which Determines the Growth of AIDS? It has been argued that the high new-partner rate among homosexuals has been the primary risk factor governing the growth of AIDS. Here we point out that if infectiousness is low, $i \ll 1$, then sexual-contact frequency rather than new-partner rate is the determining risk factor, provided the new-partner rate is greater than zero. First we note that an infected individual must infect *on the average* just one previously uninfected individual within the doubling time to produce a doubling of the number of cases. Since the doubling time of the infection has always been long compared to the new-partner exchange time (the current doubling time of the infection is more than 2 years), it is difficult to see how new-partner rate per se can be the primary risk factor. More partners and fewer sexual contacts per partner within the doubling time should transfer infection at the same rate as fewer (but some) new partners and more sexual contacts per partner. The most likely case is that new-partner rate and sexual-contact frequency are strongly correlated, but the available data are inadequate to confirm that hypothesis.

The observed correlation between high new-partner rate and infection could also be explained by the existence of a short period (several days to a week) of very high infectiousness (≈ 1) soon after initial infection followed by a long period (about 2 years) of low infectiousness. During a highly infectious period of such short duration, a victim of transfusion-related AIDS is not likely to infect his or her partner, but a homosexual with a high new-partner rate is. Thus a short spike of very high infectiousness is consistent with the high initial growth rate of AIDS (a doubling time of less than 6 months) observed among high-risk homosexuals and intravenous-drug users and with the long time (an average of more than 3 years) required for transfusion-infected people to infect their spouses.

Later we will discuss the role that a variability in infectiousness from person to person might play in the question of whether new-partner rate or sexual-contact frequency governs the growth rate of the epidemic. In any case, whichever is the causative risk, both can be described by an inverse cubic distribution, provided we

assume infectiousness is not correlated with risk behavior. Thus we assume the following distribution of risk behavior:

$$\frac{N(r)}{N_0} = \begin{cases} 1 & \text{for } r < 1 \\ r^{-3} & \text{for } r \geq 1, \end{cases} \tag{13}$$

where r is normalized so that a value of 1 is assigned to the mean value of r and $\frac{3}{2}N_0$ is the size of the population.

The Saturation Wave

We are now in a position to describe how the infection travels through the population. We start with our assumption of biased mixing, namely, that the population is divided into groups of individuals with similar risk behavior and that the members of each group interact primarily among themselves (intragroup preference). Since the relative growth rate of the infection is proportional to the risk behavior r, the time for the epidemic to approach saturation within each group will be proportional to r^{-1}. Also, higher-risk groups have fewer members than lower-risk groups, so higher-risk groups saturate much faster than lower-risk groups. Thus, after a member of the highest risk group is infected, that group quickly saturates, then the next lower group saturates, and so on. We say that a "saturation" wave of infection travels from high- to low-risk groups.

SATURATION WAVE PRODUCED BY BIASED MIXING

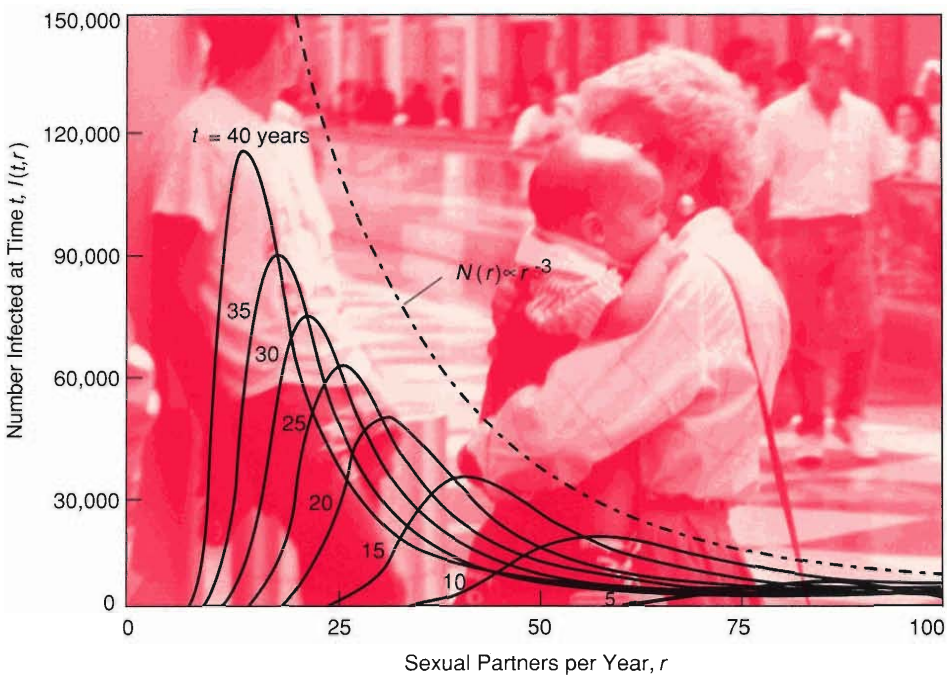


Fig. 7a. When the mixing among a population is biased (that is, when individuals with similar risk behavior (here new-partner rate) interact primarily among themselves), our model predicts the distributions by risk behavior of the number infected shown on the right. The distributions were calculated for various times t after an individual with very high risk behavior became infected. Note that the number infected approaches saturation first in the highest-risk group and then, as time passes, in successively lower and lower risk groups. We describe that situation by saying that a wave of saturation travels from high- to low-risk groups. Also shown (dashed line) is the initial distribution by risk $N(r)$ of the population, which is assumed to be an inverse cubic distribution.

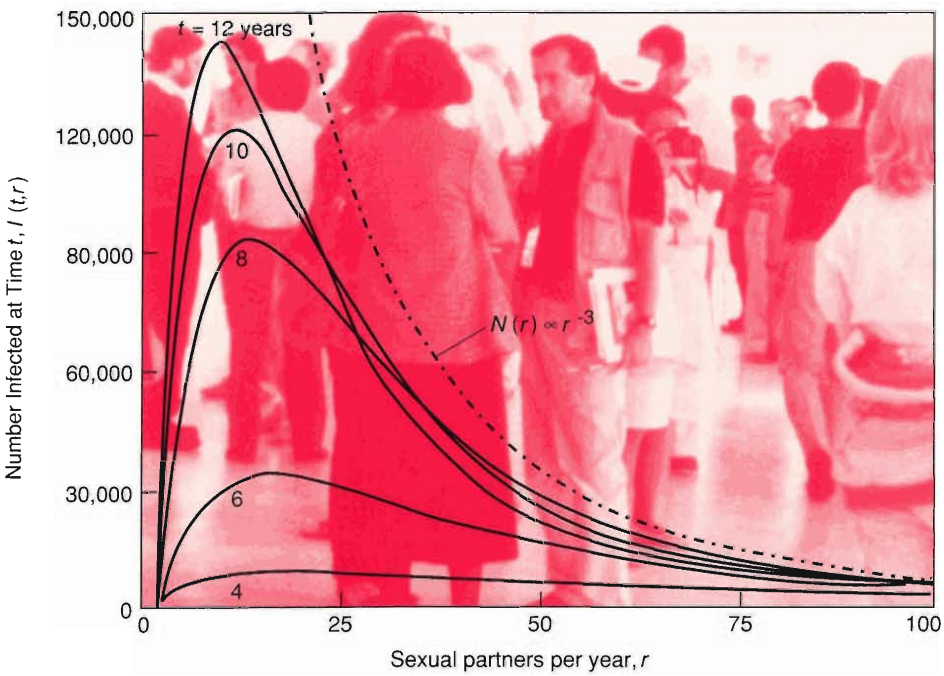
Figure 7a shows "snapshots" of the saturation wave of infection at successive times, calculated numerically from our general model. Note that the calculation separates those who progress to AIDS and death from those who are infected but do not yet have AIDS. Consequently, the plots of number infected versus risk value in Fig. 7a are always below the dotted curve representing the original distribution of risk among the population.

As time progresses, the wavefront (the low-risk end of each curve) moves from right to left, that is, from higher to lower risk values. At any given time all groups with risk values to the right of the wavefront are saturated, and all groups with risk

values to the left of the front have just a few infected members. It is primarily within the group composing the wavefront that the multiplication is taking place, and therefore the doubling time of the epidemic at any given time is primarily the doubling time of that group. Since within that active group the mixing is homogeneous, the number of infected within the group is growing exponentially, and, on average, each infected individual infects only one person within the group's doubling time.

The general model used to calculate the wave in Fig. 7a allows a small amount of mixing between groups. In addition, we allow for the possibility that some individuals were "seeded," or infected, before the start of the saturation wave. Therefore, the numbers of infected in all groups with risk values to the left of the wavefront are also growing exponentially, but at a relatively slow rate, and all groups with higher risk values are saturated, exhibiting no further growth in numbers of infected. Only the total number of infected individuals (the sum of the infected in all groups) is growing as a power law.

Fig. 7b shows what happens when we assume homogeneous rather than biased mixing. Note that the saturation wave moving from high- to low-risk groups disappears. Instead, the number infected in the average-risk group is always larger than the number infected in high-risk groups. Thus homogeneous mixing contradicts the finding of the CDC that most early victims of AIDS were high-risk individuals. Moreover, homogeneous mixing yields exponential rather than power-law growth.



NO SATURATION WAVE WITH HOMOGENEOUS MIXING

Fig. 7b. When the mixing among a population is homogeneous rather than biased, the saturation wave in Fig. 7a disappears. Instead the maxima in the distributions of number infected always occur in low-risk groups, even early in the epidemic. Such a situation is contrary to the findings of the Centers for Disease Control.

Calculation of the Saturation Wave. We will now make the above qualitative description of the saturation wave into a quantitative model. For simplicity we ignore intergroup mixing and calculate the wave of infection as if each risk group grows independently to saturation. However, such a simplistic calculation yields essentially the same results as the more complete model that includes a small amount of mixing between groups (see "Numerical Results of the Risk-Based Model of AIDS").

Once the saturation wave starts, the total number of infected I at any given time is roughly the sum of all individuals from the highest-risk individual down to individuals with risk behavior r_*, the value of r at the front of the saturation wave. Thus the number of infected is equal to the integral of all individuals with $r \geq r_*$:

$$I(r_*) = \int_{r_*}^{\infty} N(r)dr = \frac{1}{2}N_0 r_*^{-2}, \tag{14}$$

where r_* is the risk behavior of the lowest risk group in which most members are infected and $N(r)$ is the number of individuals with risk behavior r, as defined in Eq. 13.

We will now convert Eq. 14 into an equation for the number of infected as a function of time. We do this by calculating the time required to saturate the group of $N(r_*)$ individuals at the front of the wave. We assume all risk groups are seeded before the start of the saturation wave at $t_* = 0$. Within each risk group the mixing is homogeneous, so the number infected with risk behavior r_* grows exponentially, or as $I(0, r_*)e^{\alpha' r_* t_*}$, where $I(0, r_*)$ is the number infected with risk behavior r_* at $t_* = 0$. Although the relative growth rate decreases as the group approaches saturation, we neglect this slowing down and say that exponential growth continues until the number infected is approximately equal to the total number in the group. (We also ignore the slow depletion of number infected by death.) Thus

$$N(r_*) \approx I(0, r_*)e^{\alpha' r_* t_*}. \tag{15}$$

Then t_* is the time to saturate the group with risk behavior r_*. Solving Eq. 15 for t_* gives

$$t_* \approx \frac{1}{\alpha' r_*} \ln\left(\frac{N(r_*)}{I(0, r_*)}\right). \tag{16}$$

To the accuracy of this model, we will consider the fraction of each group initially infected, $N(r_*)/I(0, r_*)$, to be slowly varying. Then Eq. 16 says that the time t_* to saturate a group with risk r_* is proportional to $1/r_*$.

We can now express Eq. 14 in terms of t_* by replacing r_* with a constant times $1/t_*$. Thus we determine that the dominant time-dependent behavior of the number infected is

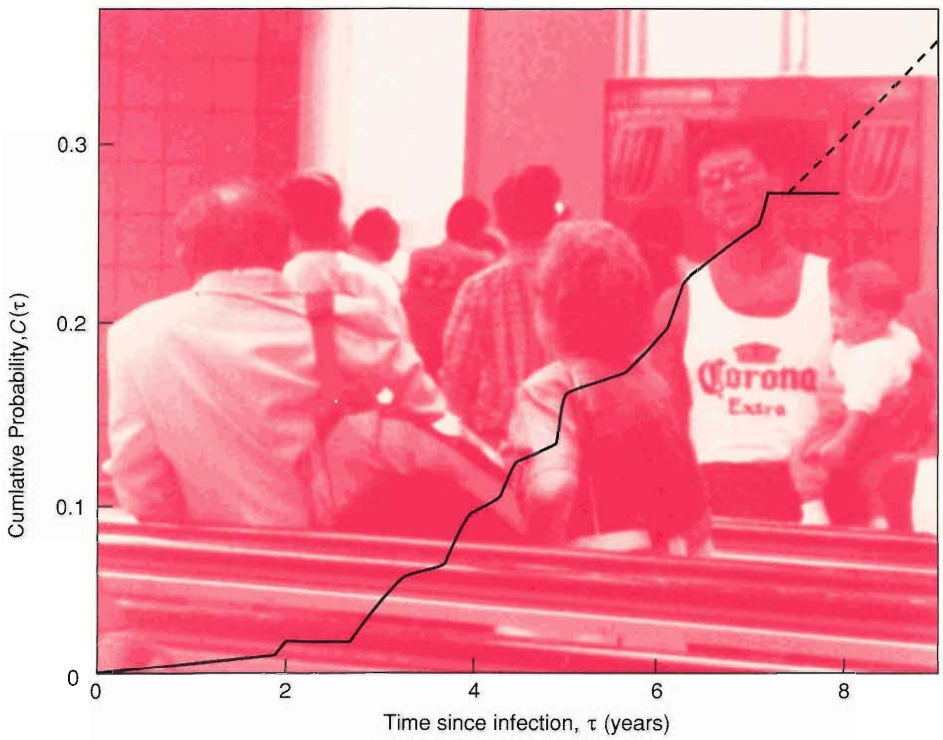
$$I(t_*) \approx I_1 t_*^2,$$

where the value of I_1 is not yet determined. We have assumed that some individuals were seeded, or infected, before a member of the highest-risk group started the saturation wave, so we add an unknown constant I_0 to obtain

$$I(t_*) \approx I_0 + I_1 t_*^2. \tag{17}$$

Although Eq. 17 cannot be valid at $t_* = 0$ (it implies that $\frac{dI}{dt_*} = 0$ at $t_* = 0$, which does not make sense), we will not attempt to refine it but instead lump all the uncertainties about the very early growth in the constant I_0. Thus we say that after the start of the saturation wave at $t_* = 0$, the number infected grows as the square of time. Since in our model the quadratic growth term $I_1 t_*^2$ will be associated with the cubic growth of AIDS, the unknown number of additional infected persons I_0 will be associated with deviations from purely cubic growth of AIDS before 1982.5.

The Progression to AIDS from Infection. Given the number infected as a function of time, we now need to estimate the resulting growth in the number of AIDS cases. The most extensive data on the conversion from HIV infection to AIDS have their origin in a study by the San Francisco Department of Health on the spread of hepatitis B among a group of homosexual men. That study took place between 1978 and 1982 and was extended in 1984 to track HIV infection. A subset of the original group continues to be monitored for clinical evidence of AIDS. The blood samples from that study have been an invaluable source for determining the time lapse be-



PROBABILITY OF DEVELOPING AIDS

Fig. 8. The cumulative probability of developing AIDS at time τ after infection $C(\tau)$ versus τ. The time of infection is assumed to be the time at which antibodies to HIV are first detected in the blood. The data were supplied by the San Francisco Department of Health. $C(\tau)$ is near zero for the first two years after infection and then increases appoximately linearly at a rate of 0.06 per year. Linear extrapolation of the data at that constant rate (dashed line) indicates that $C(\tau) = 0.5$ at 10 years after infection and $C(\tau) \approx 1$ at 18 years after infection. Recent data extending to 10 years after infection agree with that extrapolation.

tween infection with HIV and the onset of AIDS.

Let $C(\tau)$ be the cumulative probability of conversion to AIDS at τ years after infection. Figure 8 is a graph of $C(\tau)$ versus τ derived from the San Francisco study for $\tau \leq 8$ years. For the first two years after infection, $C(\tau)$ is nearly zero. Then it increases almost linearly at a rate of 0.06 per year. Newly gathered data extend the steady rise to 10 years after infection.

The apparently inexorable increase in $C(\tau)$ is consistent with the steady decline, with time since infection, in the number of T4 lymphocytes in the blood of infected persons. Those so-called T4 helper cells are central players in the functioning of the immune system, and their demise results in a progressively decreasing ability of the immune system to destroy invading pathogens. Moreover, the rate of T4 cell destruction found in an infected person is correlated with the time required for that person to convert to AIDS. These facts suggest that HIV infection always proceeds to AIDS, as does a study by Bordt et al. of infected individuals in Frankfurt, East Germany. More than 90 per cent of that study group progressed from one stage of immune destruction to the next. The Frankfurt data indicate that at least 90 per cent of those infected will develop AIDS. Thus, even though the San Francisco study covers only 10 years of experience, we argue that a reasonable extrapolation of the data is to assume a constant rate of change in cumulative conversion probability of 0.06 per year starting 2 years after infection. In other words, we assume that

$$\frac{dC(\tau)}{d\tau} = \begin{cases} 0 & \text{for } 0 \leq \tau \leq 2 \\ 0.06 & \text{per year for } 2 < \tau < 18 \\ 0 & \text{for } \tau \geq 18. \end{cases} \qquad (18)$$

Equation 18 implies that the cumulative probability of converting to AIDS is 50 per cent at 10 years after infection and 100 per cent at 18 years after infection.

Because conversion to AIDS has a nonzero probability of happening at any time between 2 to 18 years after infection, the growth rate in the number of AIDS cases, $dA(t)/dt$, at any given t is the sum over past times τ of the product of the growth rate of newly infected at $t - \tau$ years, $dI(t - \tau)/dt$, and the differential probability

of conversion to AIDS at τ years since infection (or at time t), which is $dC(\tau)/d\tau$. That complicated sum is written as a convolution integral over past times τ:

$$\frac{dA(t)}{dt} = \int_0^\infty \frac{dI(t-\tau)}{dt}\frac{dC(\tau)}{d\tau}d\tau. \tag{19}$$

Using Eq. 18, we reduce Eq. 19 to

$$\frac{dA(t)}{dt} = 0.06\int_2^{18}\frac{dI(t-\tau)}{dt}d\tau, = 0.06[I(t-2) - I(t-18)] \text{ for } t \leq 18 \text{ years.} \tag{20}$$

Replacing $I(t-2)$ with Eq. 17, neglecting $I(t-18)$ because it is small, and evaluating $\frac{dA}{dt}$ from Eq. 2, we obtain

$$3A_1t^2 = 0.06[I_1(t_* - 2)^2 + I_0] \text{ for } t > 1.3 \text{ years.} \tag{21}$$

Thus we see that if

$$I_1 = \frac{3A_1}{0.06} \approx 8700, \tag{22}$$

$$t_* = t + 2, \tag{23}$$

and if I_0 is small compared to $(1.3)^2I_1 \approx 15,000$, then our model fits very closely the AIDS case data in Eqs. 1 and 2. The time shift of 2 years reflects our approximation that AIDS does not develop during the first two years following infection.

Equation 17 for the number of infected becomes

$$I(t) = 8700(t+2)^2 + I_0, \tag{24}$$

where t is the time since 1981.2. This equation will be valid from the start of the saturation wave, which occurs before $t = 1.3 - 2$ years $= -0.7$ years (1980.5). Hence we estimate that in 1988.2, or $t = 7$ years, the number of infected persons that will eventually be reported as CDC-defined AIDS cases (using the pre-1987.5 definition) was

$$I \approx 8700(9)^2 \approx 700,000. \tag{25}$$

To summarize, our biased-mixing, risk-based model shows a cubic growth of AIDS independent of learning and predicts that the infected population initially grew as the square of time. Both the number infected and the number of AIDS cases have doubling times that increase linearly with time. We have associated the cubic growth in AIDS cases with a quadratic growth in infections, which is produced by a saturation wave moving from high- to low-risk groups. We have not discussed what happens prior to the start of the saturation wave, since that is more speculative. However, in "The Seeding Wave" we present a plausible scenario for the initial spreading of infection.

Consequences of the Model

We use the simple model described above to answer a number of questions. These questions are also relevant to our general model and to other more complex models still to be developed.

Present Number Infected. The estimate of about 700,000 infected in 1988.2 is significantly less than the estimate of 1.5 million made several years ago but agrees more closely with the CDC estimates of 1 to 1.5 million. While the earlier num-

ber was probably an overestimate, the estimate obtained from Eq. 24 was not corrected for cases not reported, which amount to about 10 per cent of the total, and for cases falling outside the pre-1987 CDC definition, which amount to about 20 per cent of the total. The estimate of 700,000 must therefore be multiplied by a factor of $1/(0.9 \times 0.8) = 1.4$. Thus our model predicts that approximately 1 million individuals in the United States were infected with HIV by 1988.2. This prediction is based on the assumption that behavioral changes due to learning did not greatly reduce the growth of infection. If learning has been effective, the number infected could be less. More likely, however, is that infectiousness depends on the stage of the disease, which, in turn, implies a greater number infected (see below).

Average Time Since Infection. To determine whether behavioral changes could have affected the growth of AIDS, we must first determine how long ago, on average, those persons now developing AIDS were infected and then question whether learning was a significant factor at that time. The mean time since infection $\bar{t}$ of the AIDS cases at time t is given by

$$\bar{t} = \frac{\int_0^\infty \frac{dI(t_*-\tau)}{dt_*}\tau\frac{dC(\tau)}{d\tau}d\tau}{\int_0^\infty \frac{dI(t_*-\tau)}{dt_*}\frac{dC(\tau)}{d\tau}d\tau} = \frac{1}{3}\frac{t_*^3 - 12t_* + 16}{(t_*-2)^2} \approx \frac{1}{3}t + 2 \text{ years.} \quad (26)$$

For 1988.2, $t = 7$ years and $\bar{t} \approx 4.3$ years. That is, those persons developing AIDS in 1988.2 became infected, on average, in 1983.9. One might expect the mean time since infection to be closer to 10 years, the time when the cumulative probability for conversion to AIDS $C(\tau)$ equals 0.5. The mean time is much shorter than 10 years because the fast growth rate of the infected population relative to the slow rate of conversion to AIDS biases the time since infection of the AIDS cases in 1988.2 closer to the time when most were infected.

Learning and Decreasing Growth Rate. We emphasize that 1983.9 is just about when learning started on a large scale, that is, when the bath houses in San Francisco were closed and safer sex practices began to be accepted. Therefore, we may expect that a decrease in the growth of AIDS among homosexual men below the cubic growth has already started. The change in the definition of AIDS makes that difficult to see in the data. Our estimate of the number infected in 1988.2 is based on extrapolating the observed initial cubic growth of AIDS cases into the future, so the actual number infected in 1988.2 may have been considerably less than a million due to learning. In any case the decreasing relative growth rate observed until early 1988 cannot be ascribed to learning.

Risk Behavior As a Function of Time. Our model suggests that individuals with the highest risk behavior are infected first and that, as time goes on, individuals with lower risk behavior become infected. We can quantify that change over time provided we have estimates of the population size and the present number infected.

We consider one sector of the population, namely, the 40 million males between the ages of 20 and 40 residing in principal American cities, and limit the group to those who actively exhibit homosexual behavior. If the Kinsey estimate for the percentage still holds, 10 per cent of the 40 million males, or 4 million, are homosexual. Equation 13 tells us that the size of the population is $\frac{3}{2}N_0$, so for the population being considered here, $N_0 = 2.7$ million. From Fig. 3 we learn that 65 per cent of the AIDS victims are homosexuals so we can equate I from Eq. 14 to $0.65I$ from Eq. 24. Neglecting I_0 we have

$$\frac{1}{2}N_0 r_*^{-2} = (0.65)8700(t+2)^2. \quad (27)$$

Substituting the value of 2.7 million for N_0 and solving Eq. 27 for r_*, we find that the risk behavior of the male-homosexual group being infected at time t varies inversely with time:

$$r_* = \left(\frac{2,700,000}{(2)(0.65)(8700)}\right)^{1/2}(t+2)^{-1} \approx 15(t+2)^{-1}. \quad (28)$$

Recall that r and r_* were normalized so that they are multiples of the average risk behavior and t is the time since 1981.2.

Thus, our model suggests, for example, that most homosexual victims of AIDS in 1988.2 were infected 4.3 years earlier when t = 2.7 years, that 200,000 were infected at that time, and that their risk behavior then was about 3 times the average behavior. More generally the model predicts that the risk behavior of those being infected is a continuously decreasing function of time and that the earliest infected, who in general were the earliest victims of AIDS, were those with the highest risk behavior. That last point coincides with the original findings of the CDC and others. In contrast, models based on homogeneous mixing (recall Fig. 7b) do not predict this time-dependent behavior, since at any time most of those being infected are members of the average- and not the higher-risk groups. The high average risk behavior at time of infection characteristic of the early cases of AIDS is a strong argument for the importance of including behavior in any model of the AIDS epidemic.

Mean Probability of Infection. We can combine results for the risk behavior as a function of time and the growth of the number infected as a function of time to estimate $\bar{i}$, the mean infectiousness, or mean probability of transferring infection per sexual contact. For example, let's consider those developing AIDS in 1988.2, who had, on average, a new-partner rate of approximately 3 times the mean.

Now suppose sexual-contact frequency is correlated with new-partner rate; that is, suppose a new-partner rate of 3 times the mean implies a sexual-contact frequency of 3 times the mean. Three times the mean sexual outlet frequency f is 450 sexual outlets per year (see Fig. 6), the major fraction of which can, according to Kinsey et al., be considered possible infectious contacts. Neglecting I_0 in Eq. 24, the relative growth rate of the infection α is given by:

$$\alpha \equiv \frac{dI/dt}{I} = \frac{2}{t+2}. \quad (29)$$

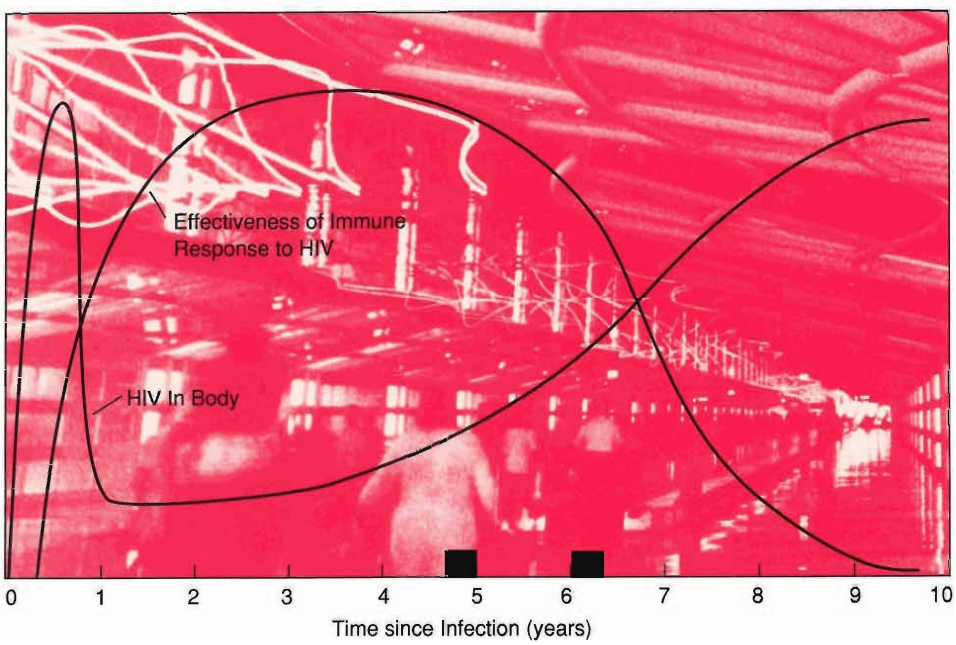
Thus at t = 2.7 years, α = 0.43 per year. Because the growth is primarily within the risk group at the front of the saturation wave, and the growth within that group is exponential, the doubling time is given by $t_d = (\ln 2/\alpha)$ = 1.6 years. On the average, each infected member of the group infects only one new partner per doubling time. Thus the average infected person has ft_d = (450)(1.6) = 720 sexual contacts and infects one previously uninfected person. In other words, $720\bar{i}$ = 1 and the mean infectiousness is approximately 0.0014. If the sexual-contact frequency is uncorrelated with new-partner rate, then we assume the sexual-contact frequency is the mean value, or 150 sexual outlets per year, and the mean infectiousness must be three times larger or about 0.004. These estimates are on the low end of the estimates of 0.003 to 0.1 made by Grant, Wiley, and Winkelstein. Also, the large uncertainties in our estimates are proportional to the uncertainties in dI/dt in 1983.9 and the uncertainties in f.

Time-Dependent Infectivity. Our estimates for the mean infectiousness (or infectivity) say nothing about the extreme variability observed from one individual to another. In an extraordinary example, four out of eight Australian women were infected with HIV from one donor sample of cryo-preserved semen split ten ways.

By contrast, in New York ninety artificial inseminations with infected semen gave rise to no infections. Is the variability due to episodic infectivity in individuals or to different strains of virus? If the Australian example were due to a particularly virulent strain of virus, several hundred times more infectious than the average, then that strain would have rapidly eclipsed all others and the growth of infection would have been many times faster since the Australian incident in 1982. Since that clearly has not happened, we must consider other possible causes of the large variability in infectivity: (1) a few individuals may be highly infectious for a period longer than the doubling time; (2) all individuals may be highly infectious for very short episodes of time; (3) highly infectious mutations may quickly mutate to less infectious ones; or (4) some individuals may be more infectious than others. Dr. G. J. Stewart has suggested that his Australian donor was in the active pre-mononucleosis-like phase of infection, which occurs before the debilitating lymphoma characteristic of pre-AIDS patients, and was therefore highly infectious. Stewart also cites three instances of infected women who have not yet (in 6 years) infected their unprotected male partners. On the other hand the very rapid spread of infection in the Kagera region of Tanzania (from only a few seropositive persons in 1984 to 43 per cent of urban adults in 1988) may indicate that a more virulent strain has emerged.

Our model tacitly assumed a constant infectivity per unit time so that the relative growth rate α was proportional to risk behavior. However data from Walter Reed Army Medical Center and other institutions suggest that the amount of virus in the blood, and therefore the infectiousness, follows the curve shown in Fig. 9. Further studies are desperately needed to pin down the course of AIDS within individuals and the resulting infectivity as a function of time, but for the moment the data shown in Fig. 9 are the best estimate we have. Those data indicate that for a brief period following infection, people are highly infectious, then for several years the immune response is able to halt viral replication, thereby reducing infectiousness to a very low level, and finally, as the immune system deteriorates and the T4 cell count declines, infectiousness rises steadily. If this pattern is correct, how does it alter the predictions of our risk-based model?

We mentioned earlier that a short period (several days to several weeks) of high infectiousness (greater than 0.5) immediately after infection could have driven the early phase of the epidemic, when new-partner rates were greater than one new part-



HIV REPLICATION AND INFECTIVITY

Fig. 9. Dependence of infectivity on time since infection most likely follows the red curve, which describes the amount of HIV in the body. Initially the virus replicates rapidly, but then the immune system mounts its defense and viral replication is stopped. At about 2 years after infection, the immune system begins to break down and viral replication resumes. The black curve depicts the effectiveness of the immune response to HIV. (The figure was adapted, with permission of Scientific American, Inc., from one appearing in the article "HIV infection: The clinical picture" by Robert R. Redfield and Donald S. Burke. *Scientific American*, October 1988.)

ner per week. A spike of high infectiousness of that duration is consistent with the observed cubic growth in the high-risk population provided it is followed by a period of low infectiousness lasting roughly several years. Further, if infectivity is correlated with decreasing T4 cell count and therefore begins rising a few years after initial infection, our model predicts a growth in AIDS cases proportional to a power of time greater than 3. Thus we expect a transition in the growth pattern of the epidemic as the saturation wave moves from groups with high new-partner rates to those with lower ones. This change would reflect the fact that among the high-risk population, the disease spreads most during the short, initial infectious period, whereas, among the low-risk population, the disease spreads most during the five to ten years of increasing infectivity in the later stages of disease. Since the heterosexual population is characterized by relatively low new-partner rates, the latter mode of growth will probably dominate in that group. (The effects of time-dependent infectivity for the more complete model are presented in "Numerical Results of the Risk-Based Model.")

Do Super Spreaders Exist? We have already pointed out that the low average infectiousness implies that sexual-contact frequency rather than new-partner rate determines the growth rate of the epidemic. However, the new-partner rate within a group can be the dominating risk factor if a small percentage of individuals within the group are highly infectious. If such individuals have more new partners but maintain the same sexual-contact frequency, they will infect more individuals. Since super spreaders infect almost every one of their partners, the fraction of such highly infectious individuals must be small to maintain the observed growth rate. The singular Australian case supports that possibility. Therefore, an understanding of the biological mechanism of high infectivity and means for identifying highly infectious individuals become important to controlling the epidemic.

High-Risk Heterosexual Groups. If a self-sustaining epidemic exists among heterosexuals, then our model suggests that it would first occur among nonmonogamous heterosexuals whose sexual-contact frequency and/or new-partner rate were several times the mean of that group or higher. At this time, a firm determination can be made only by choosing a large enough sample of those high-risk individuals and determining that more of them are infected than can by explained by unwitting contacts with homosexuals and intravenous-drug users. The experience of interviewers has shown that many people who initially claim only heterosexual risk may not be telling the truth. This creates a bias among researchers that anyone denying other risks is either lying or mistaken (for example, female contacts of intravenous-drug users may be ignorant of their partner's drug habit).

Masters, Johnson, and Kolodny have made an attempt to choose a high-risk, purely heterosexual sample by selecting heterosexuals who had more than 5 new partners per year for 5 years running. (They estimate that less than 5 per cent of the nonmonogamous, sexually active heterosexual population satisfy that criterion.) They found that 6 per cent of that group was infected. Their study has been severely criticized on methodological grounds. Although we are not in a position to defend the details of the study, we do believe that their philosophy was correct: the only way to make an early estimate of the spread of AIDS among heterosexuals is to look at the high-risk end of that population. Without such studies the disease may spread silently as behavior goes unchanged among a population that believes it is not at risk.

Conclusions

We have constructed a risk-based, biased-mixing model that reproduces the observed cubic growth of AIDS when: the risk behavior, quantified as r, is distributed

among the population as r^{-3}; either new-partner rate or sexual-contact frequency dominates the risk behavior or both are positively correlated; and the cumulative probability of conversion to AIDS increases at an approximately constant rate. The implications predicted by or consistent with the model are many. In the hope that those implications will inspire further research and promote greater awareness of the threat of AIDS, we end by listing them.

- The total number of persons infected with HIV in 1988 was roughly one million.

- The mean time between infection and onset of AIDS is an increasing function of time.

- The decreasing relative growth rate of AIDS cases observed through 1988 was not due to changes in behavior.

- The mean risk behavior of AIDS victims at time of infection is a decreasing function of time.

- The mean probability of infection per sexual contact may be as small as 0.004 to 0.001.

- A slow increase in infectivity during the progression from infection to AIDS could change the growth of AIDS from the cubic growth rate now observed to something faster, and behavior modification could change it to something lower.

- New-partner rate is the dominant risk factor if sexual-contact frequency and new-partner rate are strongly correlated or if a few per cent of the population have a very high infectiousness; otherwise sexual-contact frequency is the dominant factor.

- Most major subpopulations, both demographic and geographic, were infected by a few high-risk individuals early in the epidemic, and only small, highly socially isolated groups may remain untouched by the epidemic.

- One likely path by which the infection initially reached the high-risk groups was by an initial seeding of the average-risk population (see "The Seeding Wave"). A seeding wave then progressed from low- to high-risk groups before 1979. Simulation of such a seeding wave suggests that the first case of infection could have occurred in the average population in the late 1960s. Only somewhat less probable is occurrence of the first case of AIDS in the higher-risk groups in the late 1970s.

- After the highest-risk group is saturated (most of its members are infected), a saturation wave of infection proceeds to lower-risk groups, producing the cubic growth in AIDS cases.

- Growth of AIDS cases within the purely heterosexual, drug-free population may also be governed by a power law (most likely cubic). However only by measuring prevalence in high-risk heterosexual groups adequately isolated from other known risk groups can such a determination be made.

- More speculative is the implication that the initial spike of the time-dependent infectivity caused the initial rapid growth in the homosexual and intravenous-drug-using populations and that the gradual increase in infectivity about two years following infection may be driving a second much slower epidemic (measured in decades) among the heterosexual drug-free population. That latter mode of slow spread may be the strategy evolved by the virus to survive in equilibrium with its human hosts.

Our risk-based model is seen by many as controversial. Certainly data on sexual behavior and mixing patterns that firmly substantiate our assumptions are sadly lacking in the literature. Even more unfortunate is the difficulty in collecting data on private behavior. The singular dedication of Kinsey must be emulated on a larger de-

mographic scale with a societal consensus of the necessity for truthful answers and the guarantee of legal protection. Such data may take many years to collect, whereas urgency is needed to help us stem the spread of this deadly disease. Thus we have used the available data to develop what we feel is a likely and plausible model for the growth of the epidemic. Whether exactly right or not, the model raises questions that we cannot ignore. It also offers simple quantitative tools to estimate the size of the problem and to quantify the effectiveness of strategies aimed at minimizing the growing threat. ■

Acknowledgments

We appreciate extensive interaction with the many people who helped us initiate research on AIDS at Los Alamos. Among them are Robert Redfield, Jim Koopman, Klaus Dietz, Roy Anderson, Lisa Sattenspiel, Robert May, Meade Morgan, and the staff of the CDC, particularly Harold Jaffe and B. H. Darrow.

Further Reading

James M. Hyman and E. Ann Stanley. 1988. Using mathematical models to understand the AIDS epidemic. *Mathematical Biosciences* 90:415–473.

Robert M. May and Roy M. Anderson. 1987. Transmission dynamics of HIV infection. *Nature* 326:137–192.

R. M. Anderson, R. M. May, and A. R. McLean. 1988. Possible demographic consequences of AIDS in developing countries. *Nature* 332:228–234.

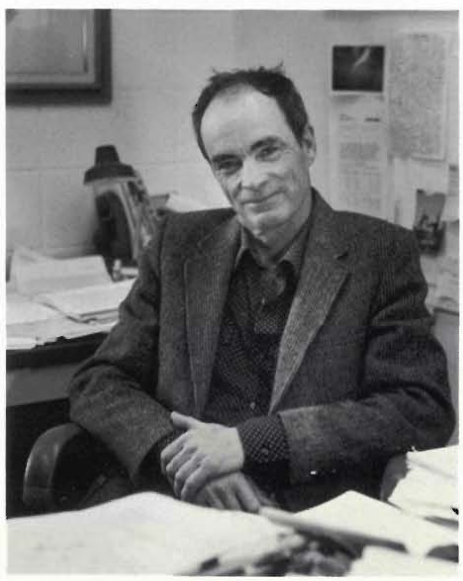
Hebert W. Hethcote and James A. Yorke. 1984. *Gonorrhea: Transmission Dynamics and Control.* Lecture Notes in Biomathematics, Volume 56. Berlin: Springer-Verlag.

Alfred C. Kinsey, Wardell B. Pomeroy, and Clyde E. Martin. 1948. *Sexual Behavior in the Human Male.* Philadelphia: W. B. Saunders Company.

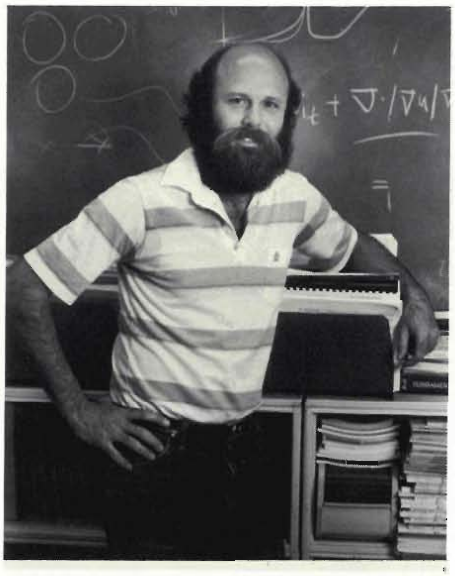
H. R. Brodt, E. B. Helm, A. Werner, A. Joetten, L. Bergmann, A. Klüver, and W. Stille. 1986. Spontanverlauf der LAV/HTLV-III-Infektion. *Deutsche Medizinische Wochenschrift* 111:1175–1180.

Robert M. Grant, James A. Wiley, and Warren Winkelstein. 1987. Infectivity of the human immunodeficiency virus: Estimates from a prospective study of homosexual men. *The Journal of Infectious Diseases* 156:189–193.

William H. Masters, Virginia E. Johnson, and Robert C. Kolodny. 1988. *Crisis: Heterosexual Behavior in the Age of AIDS.* New York: Grove Press.

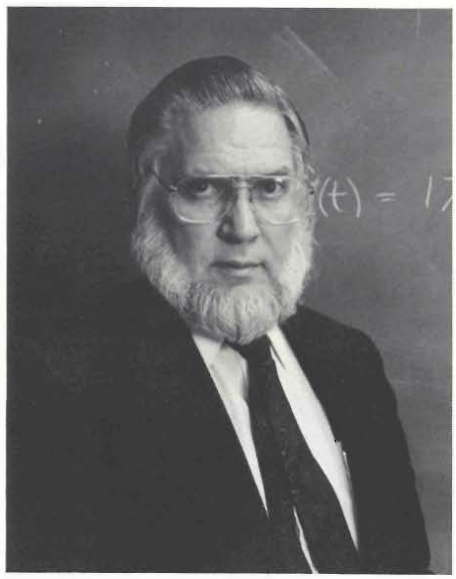


Stirling A. Colgate received his B.S. and Ph.D. degrees in physics from Cornell University in 1948 and 1952, respectively. He was a staff physicist at Lawrence Livermore Laboratory for twelve years and then president of New Mexico Institute of Mining and Technology for ten years. He remains an Adjunct Professor at that institution. In 1976 he joined the Theoretical Division at Los Alamos and in 1980 became leader of the Theoretical Astrophysics Group. In 1981 he became a Senior Fellow at the Laboratory. He is a member of the National Academy of Sciences and a board member at the Santa Fe Institute. His research interests include nuclear physics, astrophysics, plasma physics, atmospheric physics, inertial fusion, geotectonic engineering, and the epidemiology of AIDS. He has been responsible for nuclear weapons testing and design, an advisor to the U.S. State Department for nuclear testing, and a group leader in magnetic fusion. His early work on supernova led to the understanding of early neutrino emission from neutron stars—since confirmed by the supernova 1987a.



James M. Hyman is currently the Administrative Manager for the Advanced Computing Facility at Los Alamos. Before coming here in 1976, he was an instructor and research assistant at the Courant Institute of Mathematical Sciences. He received his M.S. (1974) and Ph.D. (1976) degrees in mathematics from the Courant Institute and two B.S. degrees, one in physics and the other in mathematics, both with honors, from Tulane University in 1972. His research interests include the development and analysis of numerical methods and software for the solution of partial differential equations. One goal of this work is to develop expert systems that automatically generate a computer code approximating the solution to mathematical models for, say, oil flow in a reservoir, laser fusion, or the weather. Recently his interest has turned toward mathematical models for understanding and predicting the AIDS epidemic.

E. Ann Stanley came to Los Alamos in 1984 as a postdoctoral fellow and became a staff member in the Mathematical Modeling and Analysis group in the Theoretical Division in 1987. She received a Ph.D. in applied mathematics from the California Institute of Technology in 1985 and a B.S. in engineering mathematics from the University of California in 1979. For her thesis she developed and analyzed mathematical models for Case II diffusion, the phenomenon in which a glassy polymer absorbs a fluid, a sharp front forms between the wet and dry regions, and the front moves forward at a speed proportional to time. After coming to Los Alamos, she continued working on this and other nonlinear diffusion problems. She became involved in the AIDS research partly because of previous work on a model of the diffusion of fox rabies across Europe. She enjoys playing the flute, taking modern dance classes, bicycling, skiing, and other outdoor activities.



Clifford R. Qualls is a professor of statistics at the University of New Mexico. He received a B.A. from California State College in 1961, an M.A. from University of California in 1964, and his Ph.D. from the University of California in 1967. His research interests include applied statistics, biostatistics, stochastic processes, and time series, and he supervises a computer center for the Department of Medicine at the University. He has been a visiting staff member at Los Alamos since 1975, working on statistical studies of neutral particle beams as well as the AIDS epidemic. He is currently President of the Albuquerque chapter of the American Statistical Association.

Scott P. Layne (see the biography following "The Kinetics of HIV Infectivity").

Mathematical Formalism

by James M. Hyman and E. Ann Stanley

We will build up the equations for our risk-based model of AIDS through successive modifications of the basic equation of epidemiology, the equation of mass action. Its simplest form is given by

$$\frac{dI}{dt} = \alpha I \left(1 - \frac{I}{N}\right), \tag{1}$$

where $I(t)$ is the number infected, N is the total population and α is a constant. Equation 1 describes the spread of HIV infection by random sexual contact among a sexually active population of fixed size N. As explained in the main text, if a population mixes homogeneously, this equation gives rise to an initial exponential growth in the number infected with constant relative growth rate of α.

As the number infected becomes comparable to the total population the growth rate will decrease, so we rewrite Eq. 1 to show that time dependence:

$$\frac{dI}{dt} = \lambda(t)S(t), \tag{2}$$

where $S(t) = N - I(t)$ is the number of persons susceptible to infection and $\lambda(t) = \alpha I(t)/N$. So far the only independent variable is time t and $\lambda(t)$ is the time-dependent relative growth rate of the number infected.

To describe the AIDS epidemic over long times, we must account for individuals who eventually develop AIDS and die. Thus the total population will not remain constant but will change with time. We divide the population into three sectors: the sexually active, uninfected susceptibles $S(t)$; those infected with HIV who do not have AIDS $I(t)$; and people with AIDS $A(t)$. We assume the susceptibles and the infected are sexually active (and therefore can infect others) but that those with AIDS are not. Thus the sexually active population $N(t)$ is equal to $S(t)+I(t)$. Moreover, we assume that people mature, or migrate, into the sexually active susceptible population and retire from it at a constant relative rate μ, so that in the absence of AIDS the susceptible population would remain constant at the value S_0, that is, $N(t) = S(t) = S_0$ in the absence of HIV.

We also introduce the parameter γ, the relative rate at which people who are infected develop AIDS, and δ, the relative rate at which people die from AIDS.

Now we can write down a set of rate equations for changes in $S(t), I(t)$ and $A(t)$ with time.

The rate of change in the number infected is like Eq. 2 except the right-hand side includes negative terms that account for decreases due to conversion to AIDS at a rate $\gamma I(t)$ and aging of the infected at a rate $\mu I(t)$:

$$\frac{dI(t)}{dt} = \lambda(t)S(t) - (\gamma + \mu)I(t). \tag{3}$$

The number of uninfected susceptibles increases through maturation of "juveniles" at a rate μS_0 and decreases through aging at a rate $\mu S(t)$ and through infection with HIV at a rate $\lambda(t)S(t)$:

$$\frac{dS(t)}{dt} = \mu(S_0 - S(t)) - \lambda(t)S(t). \tag{4}$$

The number of people with AIDS increases through conversion of infecteds at a rate $\gamma I(t)$ and decreases through death at a rate $\delta A(t)$:

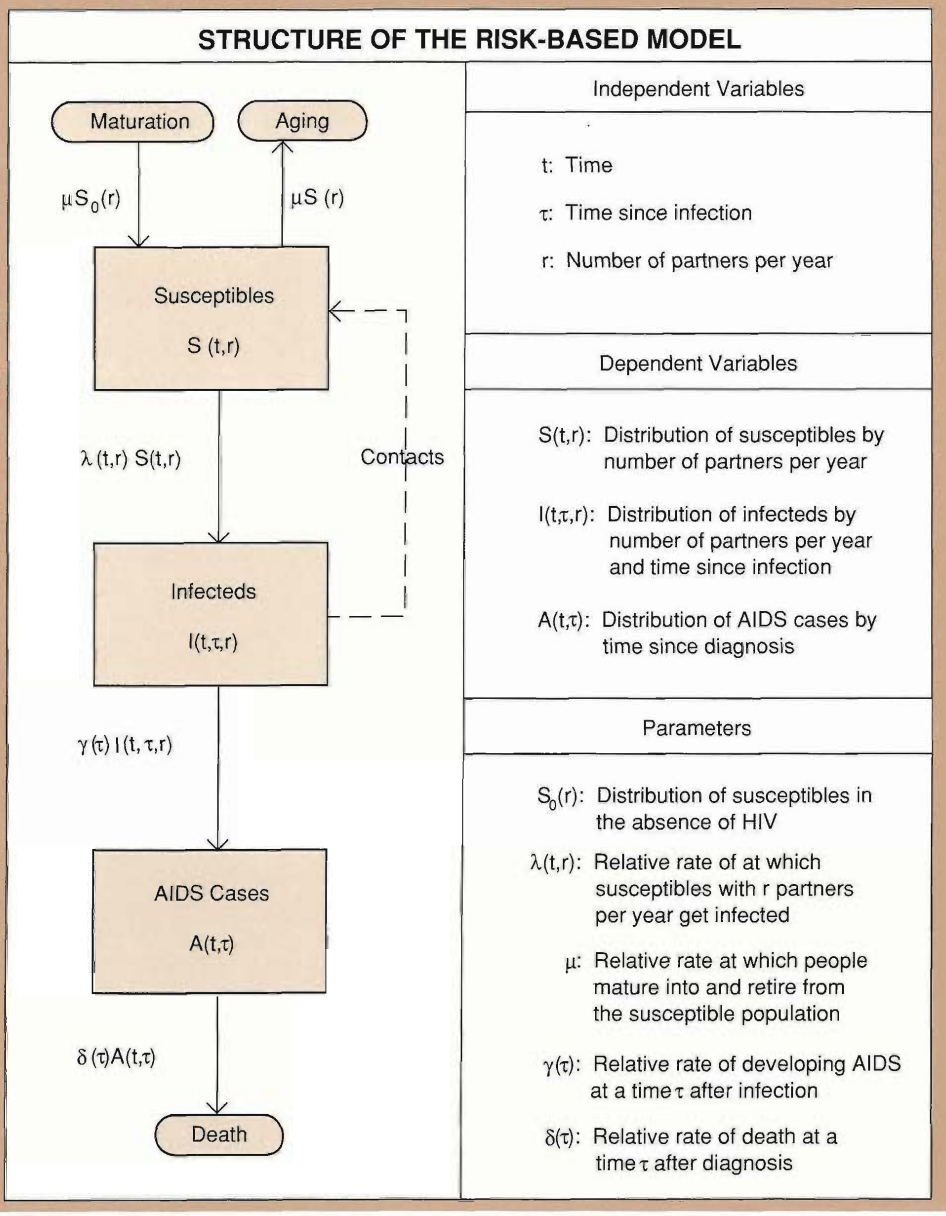
$$\frac{dA(t)}{dt} = \gamma I(t) - \delta A(t). \tag{5}$$

The accompanying block diagram illustrates the inputs and outputs to each of the three sectors of the population.

The most important assumptions in any model of AIDS are embedded in the definition of $\lambda(t)$, the rate of infection per susceptible. In the simple model just presented, all members of the population are assumed to be equal in their susceptibility and the rate of infection per susceptible is given by

$$\lambda(t) = i\ c\ p\ \frac{I(t)}{I(t)+S(t)}, \tag{6}$$

where the constant i is the probability of infection per sexual contact, the constant c is the average number of sexual contacts per partner, the constant p is the average number of partners per year, and $\frac{I(t)}{I(t)+S(t)}$ is the infected fraction of the sexually active population.



Note that this simple model produces exponential growth at the start of the epidemic. All members are equally at risk (homogeneous mixing) and the probability of infection per contact i remains constant throughout the years of infection.

We will now modify the simple model defined by Eqs. 3–6 to account for two crucial aspects of the AIDS epidemic. First, since AIDS takes many years to develop and the infectivity during the period of infection may vary in time, we introduce an additional independent variable τ, the time since infection. Second, since individuals who are very active sexually and who change partners frequently have a greater risk of becoming infected, we introduce the variable r, which quantifies the level of risky behavior in the sexually active population. In this model, r is defined as the number of new partners per year.

Using the two new independent variables τ and r, we distribute $I(t), S(t)$ and $A(t)$ over risk behavior and/or time since infection. (See the definitions in the block diagram.) In addition, the constant S_0 is the integral of an equilibrium distribution over risk behavior, $S_0 = \int_0^\infty S_0(r)dr$. Note that $S_0(r)$ corresponds to $N(r)$ in the main text; also the main text presents evidence that $S_0(r) \propto r^{-3}$ for large r.

We can now write down the equations of our risk-based model that correspond to Eqs. 3–5. Equation 3 for the infected population is replaced by Eqs. 7a and b. Equation 7a specifies that the rate at which people of risk r are becoming infected is $\lambda(t,r)S(t,r)$. Equation 7b says that rate at which the infecteds develop AIDS is proportional to the conditional probability $\gamma(\tau)$, which is a function of the time since infection, and the rate at which they leave the population is proportional to μ.

$$I(t,0,r) = \lambda(t,r)S(t,r). \tag{7a}$$

$$\frac{\partial I(t,\tau,r)}{\partial t} + \frac{\partial I(t,\tau,r)}{\partial \tau} = -\gamma(\tau)I(t,\tau,r) - \mu I(t,\tau,r). \tag{7b}$$

Equation 8 for the susceptibles has a structure similar to that of Eq. 4 except that now the rate of infection per susceptible $\lambda(t,r)$ depends on the risk behavior r:

$$\frac{\partial S(t,r)}{\partial t} = \mu(S_0(r) - S(t,r)) - \lambda(t,r)S(t,r). \tag{8}$$

Equation 9a says that the rate at which AIDS cases are being diagnosed at time t is equal to the rate at which infecteds convert to AIDS, $\gamma(\tau)I(t,\tau,r)$, integrated over all risk behaviors r and times since infection τ. Equation 9b accounts for loss of AIDS cases due to death.

$$A(t,0) = \int_0^\infty \int_0^\infty \gamma(\tau)I(t,\tau,r)d\tau dr. \tag{9a}$$

$$\frac{\partial A(t,\tau)}{\partial t} + \frac{\partial A(t,\tau)}{\partial \tau} = -\delta(\tau)A(t,\tau). \tag{9b}$$

The major change in this new set of equations is the form we assume for $\lambda(t,r)$, the relative rate at which susceptibles with r partners per year get infected. We generalize Eq. 6 to include variation in the degree of sexual contact between individuals with different risk behaviors as well as variation in infectiousness with time since infection. The general form of $\lambda(t,r)$ is given by

$$\lambda(t,r) = r \int_0^\infty \int_0^\infty c(r,s)\rho(t,r,s)i(\tau)\frac{I(t,\tau,s)}{N(t,s)}d\tau ds, \tag{10}$$

where $c(r,s)$ is the average number of sexual contacts in a partnership between a person with risk r and one with risk s, $i(\tau)$ is the infectiousness at τ years since in-

fection, $\frac{I(t,\tau,s)}{N(t,s)}$ is the probability that a person with risk s will be infected at time τ, and $\rho(t,r,s)$ is the fraction of the partners of a person with risk r who have risk s. The total number of sexually active people with risk s is given by $N(t,s) = S(t,s) + \int_0^\infty I(t,\tau,s)d\tau$.

Equations 7–10 describe the basic structure of our risk-based model. It differs from the well-known model of Anderson and May in one major respect—the form of $\lambda(t,r)$. Anderson and May assumed homogeneous mixing among the entire population, that is, that partners are chosen purely on the basis of availability. Then $\rho(t,r,s)$, the fraction of the partners of a person with risk r who have risk s, is just the proportionate mixing value:

$$\rho(t,r,s) = \frac{sN(t,s)}{\int_0^\infty xN(t,x)dx}. \tag{11}$$

They also assumed that the average number of sexual contacts per partner and the infectiousness were constant, so that $\lambda(t,r)$ becomes

$$\lambda(t,r) = \frac{icr\int sI(t,s)ds}{\int xN(t,x)dx}. \tag{12}$$

This form for $\lambda(t,r)$ (adapted from the model of Hethcote and Yorke for the spread of gonorrhea) yields exponential growth for the early stages of the epidemic.

We suggest that the assumption of homogeneous mixing is sociologically unrealistic. Instead, we build into our model a general form for $\rho(t,r,s)$ that allows for biased mixing among the population. That is, $\rho(t,r,s)$ includes an acceptance function, $f(r,s)$, that specifies the frequency at which an individual with risk behavior r chooses a partner with risk behavior s. When the acceptance function $f(r,s)$ is 1, we return to homogeneous mixing. When $f(r,s)$ is a narrow Gaussian, for example, $f(r,s) = \exp\left(-(r-s)^2/\epsilon(r+a)^2\right)$, people choose partners who are similar to themselves. This latter assumption is presented in the main text and yields the power-law growth in AIDS cases seen in the data.

For completeness we give the general form of $\rho(t,r,s)$:

$$\rho(t,r,s) = \begin{cases} (1-\int_0^r \rho(t,r,x)dx)\frac{f(r,s)sN(t,s)}{\int_r^\infty f(r,x)xN(t,x)dx}, & \text{for } r \le s \\ \rho(t,s,r)\frac{sN(t,s)}{rN(t,r)}, & \text{for } r > s. \end{cases} \tag{13}$$

This complicated function satisfies three necessary properties:

1. The number of partners with risk behavior s chosen by people with risk behavior r is equal to the number of partners with risk behavior r chosen by people with risk behavior s; that is,

$$rN(t,r)\rho(t,r,s) = sN(t,s)\rho(t,s,r). \tag{14}$$

2. People with risk behavior r have r partners per unit time; that is,

$$1 = \int_0^\infty \rho(t,r,s)ds. \tag{15}$$

3. The fractions $\rho(t,r,s)$ are positive.

In order to study the effects of different mixing patterns on the growth of the epidemic, we have chosen various forms for the acceptance function $f(r,s)$ and then solved Eqs. 7–9 numerically. The results are presented in "Numerical Results of the Risk-Based Model of AIDS." Also presented there are numerical solutions for different assumptions about infectiousness from time since infection. ■

Numerical Results of the Risk-Based Model

by James M. Hyman, E. Ann Stanley, and Stirling A. Colgate

Here we will present numerical solutions to the full risk-based biased-mixing model. These solutions validate the simplified version of the model presented in the main text and illustrate how variations in the input parameters affect the predicted course of the epidemic. The equations and parameters of the model are defined in "Mathematical Formalism for the Risk-Based Model of AIDS," hereafter referred to as "Math Formalism." The model tracks the time evolution of three sectors of the population: the sexually active susceptibles $S(t,r)$; the sexually active infecteds $I(t,\tau,r)$; and the people with AIDS $A(t,\tau,r)$. It takes into account deaths due to AIDS and the long time between HIV infection and conversion to AIDS. It also allows us to vary assumptions about the infectiousness as a function of time since infection and the mixing between various risk groups in the population.

First we will assess the validity of the predictions in the main text. The analytic calculation presented there predicted that biased mixing among the sexually active population gives rise to a saturation wave of infection, which yields power-law growth in both the number infected and the number of people with AIDS. That calculation was based on the following assumptions: the initial susceptible population $S_0(r)$ is distributed in risk behavior as r^{-3} for r greater than the mean value of r; the infectiousness i is constant; the cumulative probability of conversion to AIDS $C(\tau)$ is zero for the first two years after infection and then increases linearly with τ at a rate such that every infected individual develops AIDS by 18 years after infection; and finally, the same fraction is infected in all risk groups before the start of the saturation wave. The wave of infection was then calculated as if each risk group had a growth rate proportional to r and grew to saturation independently of all other groups. That is, we did not account for mixing between people with different risk behavior because the calculation is too difficult to perform analytically. Moreover, AIDS cases and deaths were not removed from the infected population. The result was that the number infected grows as t^2 and the number of people with AIDS grows at t^3.

To check whether mixing among individuals with different risk behavior alters that result, we solved the full set of equations given in "Math Formalism." We used the same assumptions and conditions outlined above except that we allowed mixing between people with different risk behavior r. We found that when mixing is restricted to people whose risk behaviors are within a factor of 2 of each other, that is, the mixing is biased, a saturation wave of infection moves from high- to low-risk groups and the number infected grows as t^2, as predicted by the analytic calculation in the main text. Also, when mixing is random, or homogeneous, that is, is based only on availability, the number infected grows exponentially, the relative growth rate is constant, and the fastest growth occurs in the population with the most likely risk. Thus, doubling times for biased mixing are shorter initially and later become longer than those for random mixing.

Now let's consider numerical solutions to the full model under more general assumptions. We will first comment on their overall behavior and then present specific solutions. The numer-

THE RATE OF INFECTION $\lambda(r,t)$

The heart of the risk-based model is the complicated functional form of the rate of infection per susceptible with risk r, $\lambda(r,t)$ (see Eqs. 10 and 13 in "Math Formalism"). We will describe this function in words:

Rate of infection for a susceptible = Number of new partners per year × Rate of sexual contact between persons with risk behaviors r and s × Infectiousness per contact × Probability that a person with risk s is infected

$$\lambda(r,t) = r\int_0^\infty\int_0^\infty c(r,s)\rho(t,r,s;f(r,s))\, i(\tau)\frac{I(t,\tau,s)}{N(t,s)}\,d\tau\,ds$$

Average number of sexual contacts in a partnership between persons with risk behaviors r and s

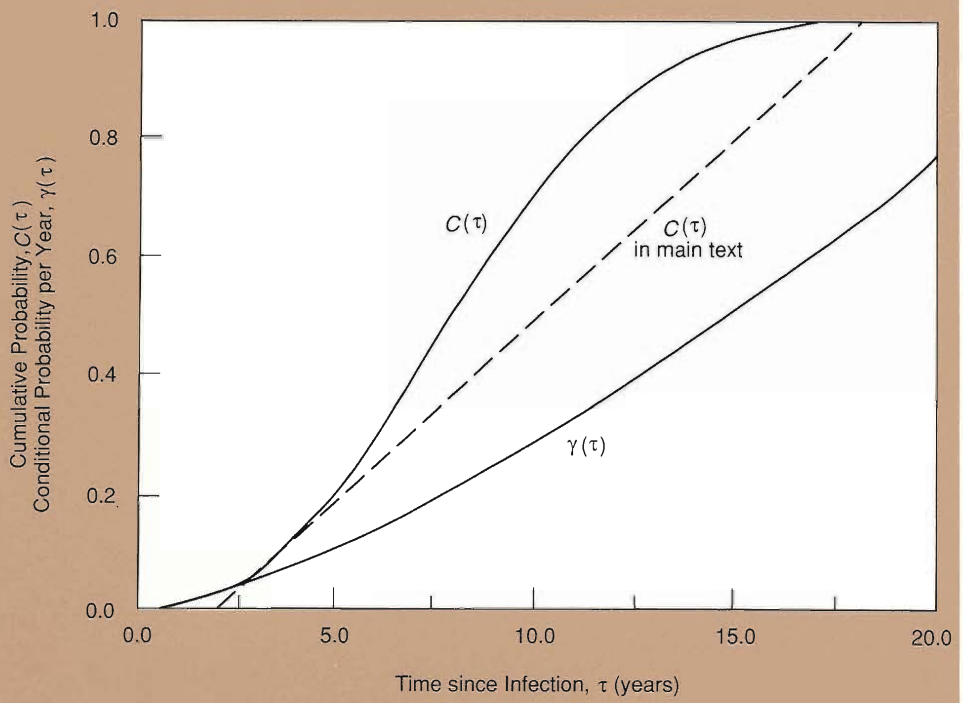
Fraction of partners of a person with risk behavior r and who have risk behavior s

The function $\rho(t,r,s)$ describes the level of mixing between people with risk behaviors r and s. It is defined in terms of an acceptance function $f(r,s)$ that determines the range from which partners are chosen.

ical results of the model change as we vary the input parameters $S_0(r)$, $I(t,0,r)$, $i(\tau)$, $A(t,0,r)$, $c(r,s)$, $f(r,s)$, $\gamma(\tau)$, μ, and $\delta(\tau)$ (see Fig. 1 in "Math Formalism" for the definitions of these parameters). The most critical parameters for determining the course of the epidemic are the initial distribution of risk behavior among the susceptible population $S_0(r)$ and the functions $i(\tau)$, $c(r,s)$, and $f(r,s)$, which determine the rate of infection per susceptible $\lambda(r,t)$ (see "The Rate of Infection"). In particular, the acceptance function $f(r,s)$ specifies the amount of mixing between different risk groups. Provided the mixing is biased, $S_0(r)$ decays as r^{-3} or r^{-4} and the numerical value of the product $c(r,s)i(\tau)$ is between 0.025 and 0.001 (this last provision determines the time scale of the epidemic), numerical solutions of our model show that the infection travels as a saturation wave from high- to low-risk groups for approximately the first 20 years. During those years the cumulative number infected and the cumulative number of people with AIDS grow as polynomials in time, rather than as exponentials.

By varying the functional forms of $\gamma(\tau)$, the rate, or conditional probability, of developing AIDS, and $i(\tau)$, the infectiousness since time of infection, we can raise or lower the degree of the polynomial growth, but in all of our calculations with biased mixing, the growth remains polynomial after the initial transients.

With these general remarks as background, we present various numerical solutions to the model. To obtain these solutions, Eqs. 9–10 in "Math Formalism" were integrated numerically with an explicit Adams-Bashford-Moulton solution method to an accuracy of 10^{-6} per unit time. The dependences on τ and r were calculated on a uniform grid of between 71 and 201 mesh points, and the convergence of solutions has been verified to within a few per cent.



RATE OF CONVERSION TO AIDS

Fig. 1. The rate of conversion to AIDS at time τ after infection $\gamma(\tau)$ is equal to the conditional probability that a person who did not have AIDS before time τ develops AIDS at time τ. Thus, it is given by $\gamma(\tau) = \frac{dC(\tau)/d\tau}{1-C(\tau)}$, where $C(\tau)$ is the cumulative probability of developing AIDS at τ years after infection. The figure shows plots of the functions $\gamma(\tau)$ and $C(\tau)$ used in all the numerical solutions presented here. For comparison we also show a plot of the form for $C(\tau)$ assumed in the main text (dashed line).

We emphasize, however, that although the solution techniques are accurate, the equations are still crude approximations and the results are meant to illustrate the general behavior of the model, not to give accurate forecasts of the future. Even the full model is much too simplistic to be used as a predictive tool.

For all the solutions presented here, we assume an initial population of 10 million people whose risk behavior (which we identify as the number of new partners per year) is distributed as an inverse cubic with a mean of 24 partners per year. We use the initial distribution $S_0(r) = 20(1 + \frac{r}{24})^{-3}$. We also use that form of $\gamma(\tau)$, the conditional probability for converting to AIDS, shown in Fig. 1. (The relationship between $\gamma(\tau)$ and $C(\tau)$ is described in the figure caption.) We use the constant value $\mu = 0.02$ per year for the fractional rate of maturation. The fractional rate of deaths due to AIDS $\delta(\tau)$ is obtained from CDC data. Also, for simplicity in this series of calculations, we assume the number of contacts per partner $c(r,s)$ is a constant $\bar{c}$.

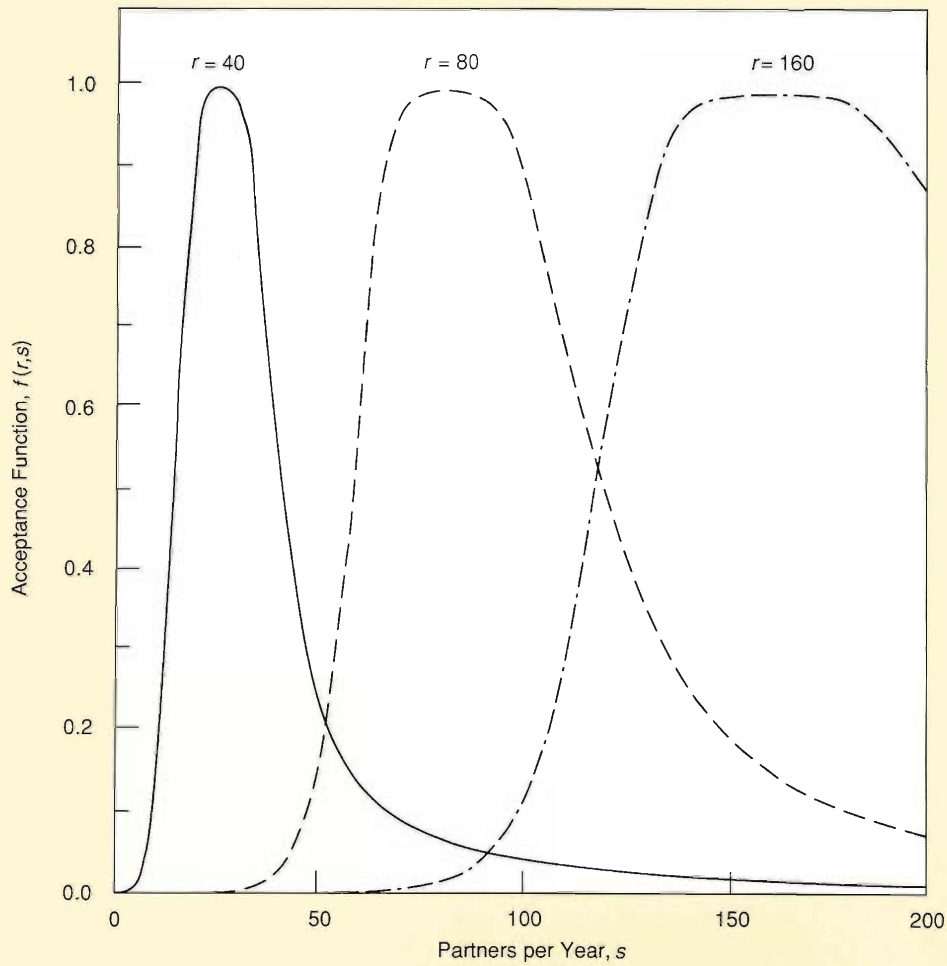
The parameter that we vary from one solution to another is $\lambda(r,t)$, the relative growth rate of infection among susceptibles with r partners per year. In particular we vary two factors in $\lambda(r,t)$: the acceptance function $f(r,s)$ and the in-

BIASED MIXING FOR BASELINE SOLUTION

Fig. 2. The numerical solutions presented here use an inverse quartic function for the acceptance function $f(r, s)$:

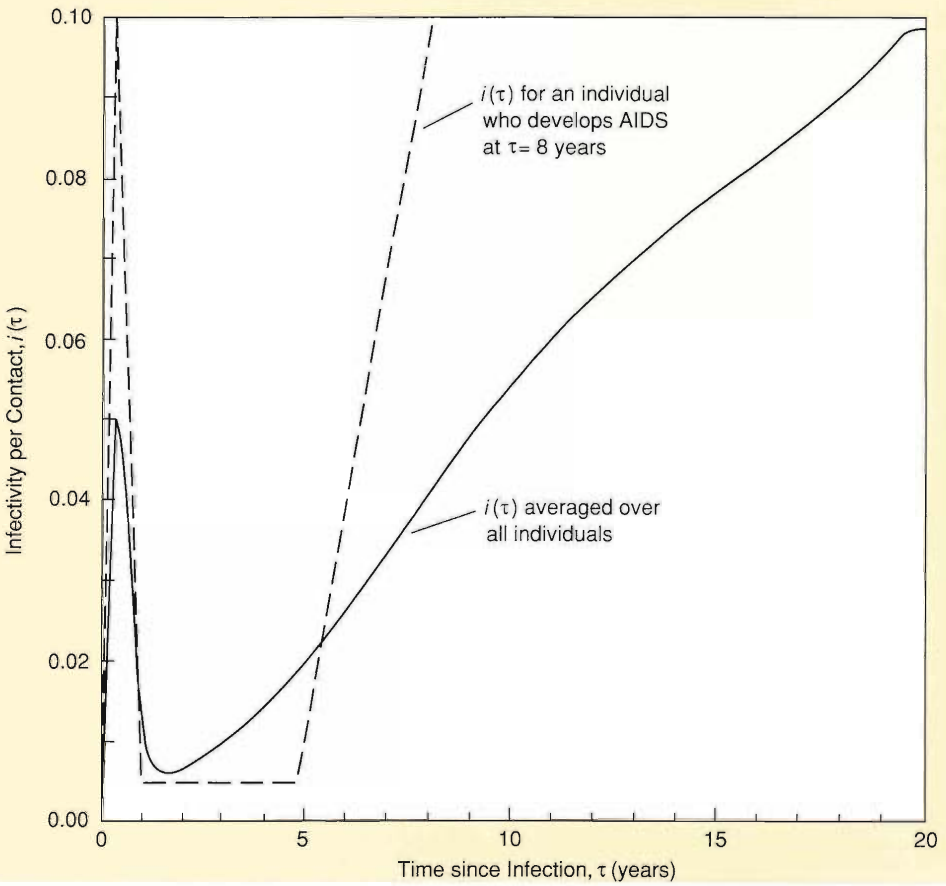
$$f(r, s) = \left[1 + \frac{(r - s)^4}{\epsilon(r + r_m)^4}\right]^{-1}.$$

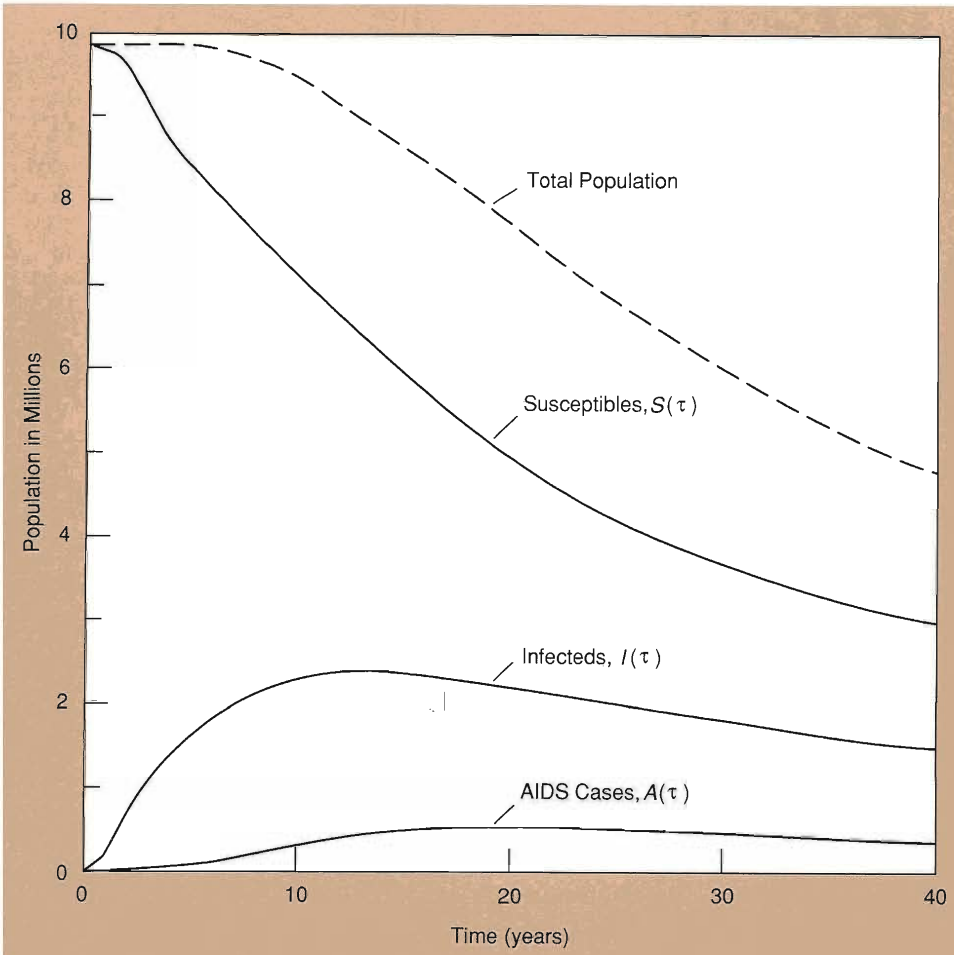
The figure shows $f(r, s)$ versus s for r = 40, 80, and 150 when ϵ = 0.01. For each value of r, $f(r, s)$ determines the fraction of partners with risk s chosen by people with risk r. Here $f(r, s)$ specifies that most partners of a person with risk r have risk behaviors between $\frac{1}{2}r$ and r; that is, the mixing is heavily biased toward people with similar risk behavior.



TIME-DEPENDENT INFECTIVITY

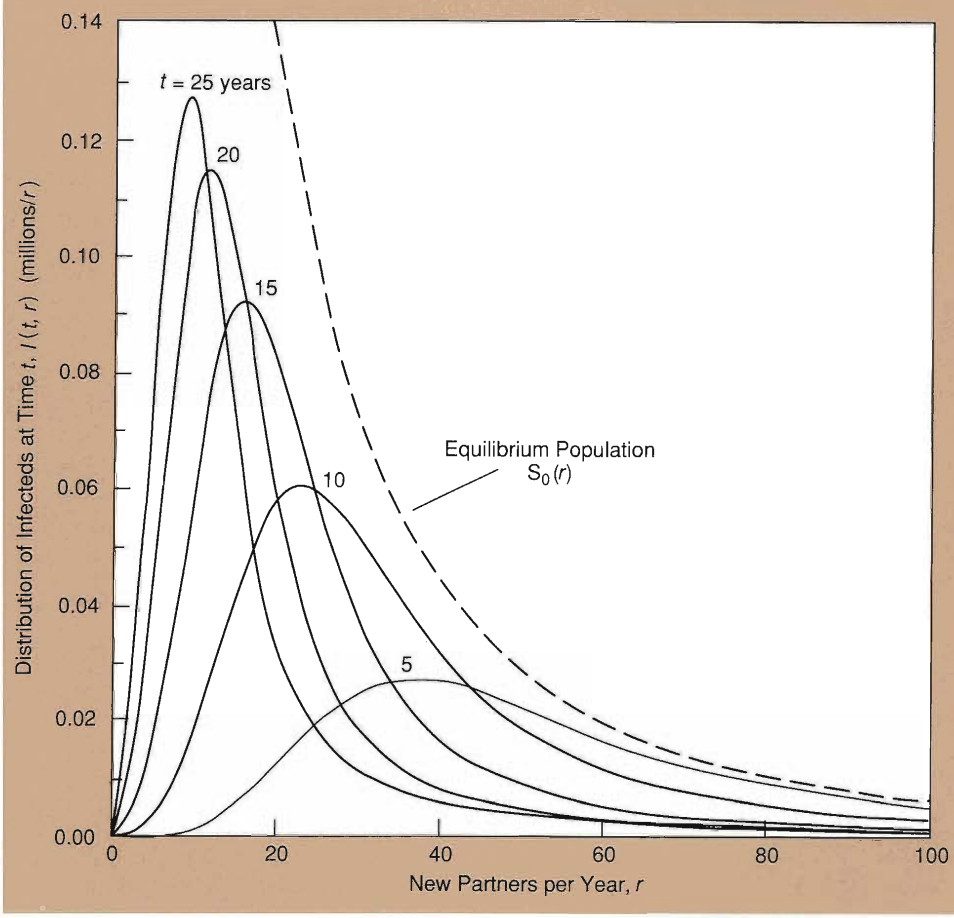
Fig. 3. The mean infectiousness $i(\tau)$ versus time since infection (solid line) used in all but the last solution presented here. The function $i(\tau)$ is an average over individuals each of whom develops AIDS at some time between 2 and 20 years since infection. The average infectiousness of each individual over the time from infection to AIDS is 0.025. The dotted line shows the pattern of infectiousness that we postulate for a single individual. In this case the individual develops AIDS 8 years after infection. The initial peak of infectiousness for this individual is always taken to be greater than 6 months because our numerical techniques are not yet designed to handle sharper peaks.





BASELINE SOLUTION

Fig. 4. The time-dependent behavior of various sectors of the population predicted by the baseline calculation. Despite a slow migration of people into the total population, the high mean new-partner rate of 24 partners per year drives an epidemic that substantially depletes the total population as a large fraction become infected and then die of AIDS. The very slow progression from infection to AIDS and rapid death from AIDS produces a delay between start of infection and the AIDS epidemic. Also, at all times many fewer people have AIDS than are infected.



SATURATION WAVE IN BASELINE SOLUTION

Fig. 5. Distributions of the number infected over number of new partners per year at times t = 5, 10, ..., 40 years during the baseline calculation. The dotted line shows the distribution of the total population in the absence of HIV. As time progresses, a wave of infection moves from high-risk to low-risk groups. Essentially all members of high-risk groups become infected, and the populations of those groups decrease to very low levels as everyone develops AIDS and dies. As the wave moves progressively through lower-risk groups, an ever smaller fraction of those groups becomes infected.

fectiousness per contact since time from infection $i(\tau)$.

We present first a "baseline" solution. The acceptance function $f(r,s)$ and the infectiousness per contact $i(\tau)$ for this solution are described in Figs. 2 and 3, respectively. The acceptance function in Fig. 2 is an inverse quartic function of r and s, which describes the probability that a person with risk behavior r chooses a partner with risk behavior s:

$$f(r,s) = \left[1 + \frac{(r-s)^4}{\epsilon(r+r_m)^4}\right]^{-1},$$

where $\epsilon = 0.01$ and $r_m = 10$ partners per year. The figure shows $f(r,s)$ versus s for three different values of r. As r increases, the width of the acceptance function increases. In rough terms, this function describes a biased mixing pattern in which a person with risk r chooses most of his or her partners from a group that ranges in risk behavior from $\frac{1}{2}r$ to $2r$.

Figure 3 is a plot of $i(\tau)$, the mean infectiousness per partnership versus time since infection. The mean infectiousness is an average over the infectiousness of many individuals each of whom develops AIDS at different times (determined by $\gamma(\tau)$) since the time of infection. Figure 3 also shows the infectiousness curve for an individual who develops AIDS 8 years after infection. The infectiousness for this individual is assumed to have an initial peak, a latency period of about four years, and finally a steady rise. The average infectiousness for each individual is assumed to be 0.025. The initial peak is about 6 months wide, probably too wide to be realistic, but our numerical code does not yet have the capability of resolving a burst that is only a few weeks in duration. Nevertheless, the wider shape that we have used serves the purpose of illustrating what the impact of an initial peak of infectiousness can be.

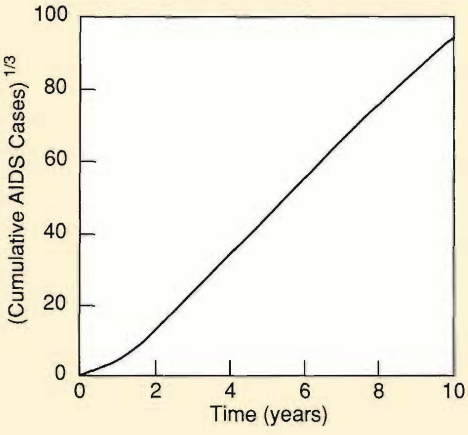
The infected population at $t = 0$ contains 1000 individuals distributed as a narrow Gaussian function of r centered at 175 partners per year and distributed linearly in τ. Although here we assume that the epidemic starts among the highest-risk groups, this choice does not have a major impact on the numerical results. In particular, if the infecteds at $t = 0$ are centered at the mean, the epidemic follows a similar course but starts about 2 years later. If the infecteds at $t = 0$ are distributed over all risk groups, the saturation wave takes off sometime between 0 and 2 years later.

The input parameters and initial conditions just described yield our "baseline" solution. Figure 4 shows $S(t), I(t)$, and $A(t)$ over a 40-year period. During that period about half of the population dies of AIDS. The number infected $I(t)$ and the number of people with AIDS at any given time $A(t)$ rise steadily for more than 10 years and then decline slightly as the epidemic reaches a steady state.

Figure 5 shows plots of the number infected versus risk behavior at times $t = 5$, 10, 15, 20 and 25 years. Here we see that the infection travels as a saturation wave from high- to low-risk groups. The wave takes 20 to 25 years to reach the lower-risk groups.

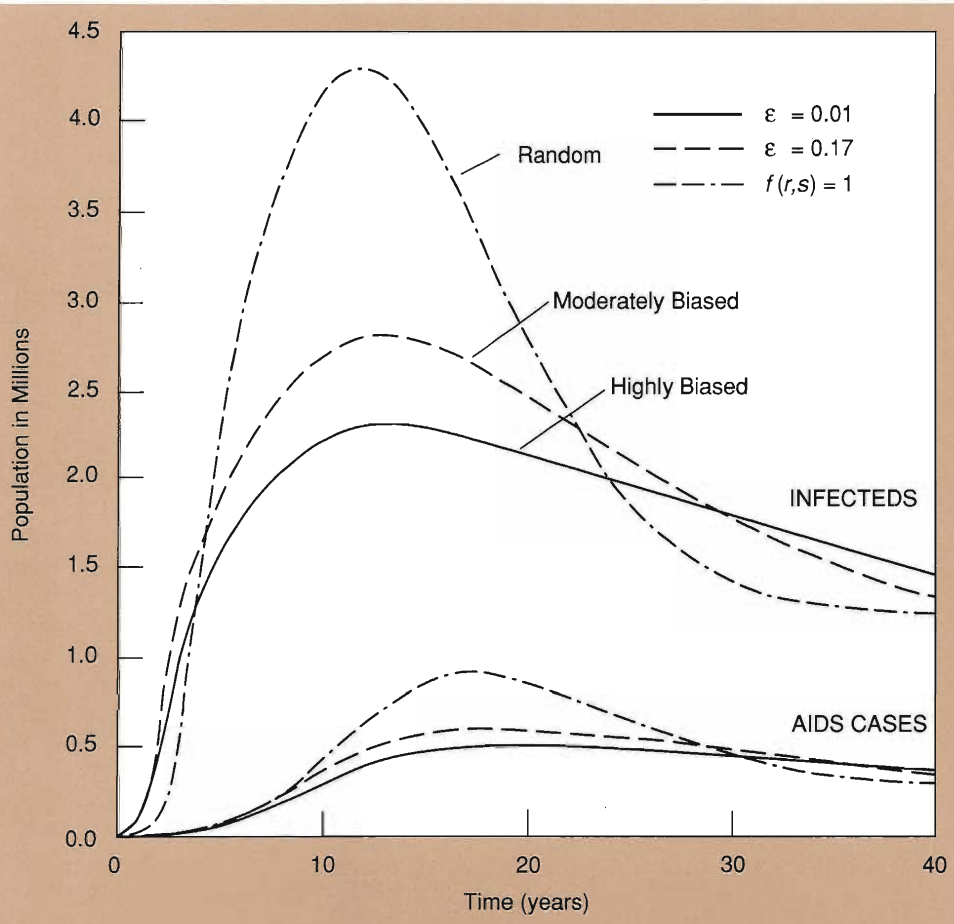
Figure 6 is a plot of the cube root of the cumulative number of AIDS cases as a function of time. The nearly straight line between 1 and 10 years shows that the calculation is not inconsistent with the observation that the number of AIDS cases grows as t^3 during the initial stages of the epidemic. The main reason that the growth is not purely cubic is the deviation of the initial profile $S_0(r)$ from a pure inverse cubic. However, the profile we chose for $S_0(r)$ fits the available partner-change-rate data much better than does Eq. 13 in the main text. We have also assumed a fairly large infectivity, which speeds up the progress of the entire epidemic. Consequently, by 10 years from the start of the saturation wave, the wave front has reached the lowest-risk populations, which, in turn, slow down the cubic growth. Although the solution just presented roughly matches the observed cubic growth of AIDS, it does not prove that the input parameters are correct but rather suggests the basic ingredients needed to produce the type of epidemic we are experiencing. A slightly different mix of input parameters yields very similar growth.

The assumption of biased mixing is the feature that sets this model apart from other models. Let's see how the epidemic changes when this assumption is relaxed. Figure 7 shows three solutions to the model that differ only in the



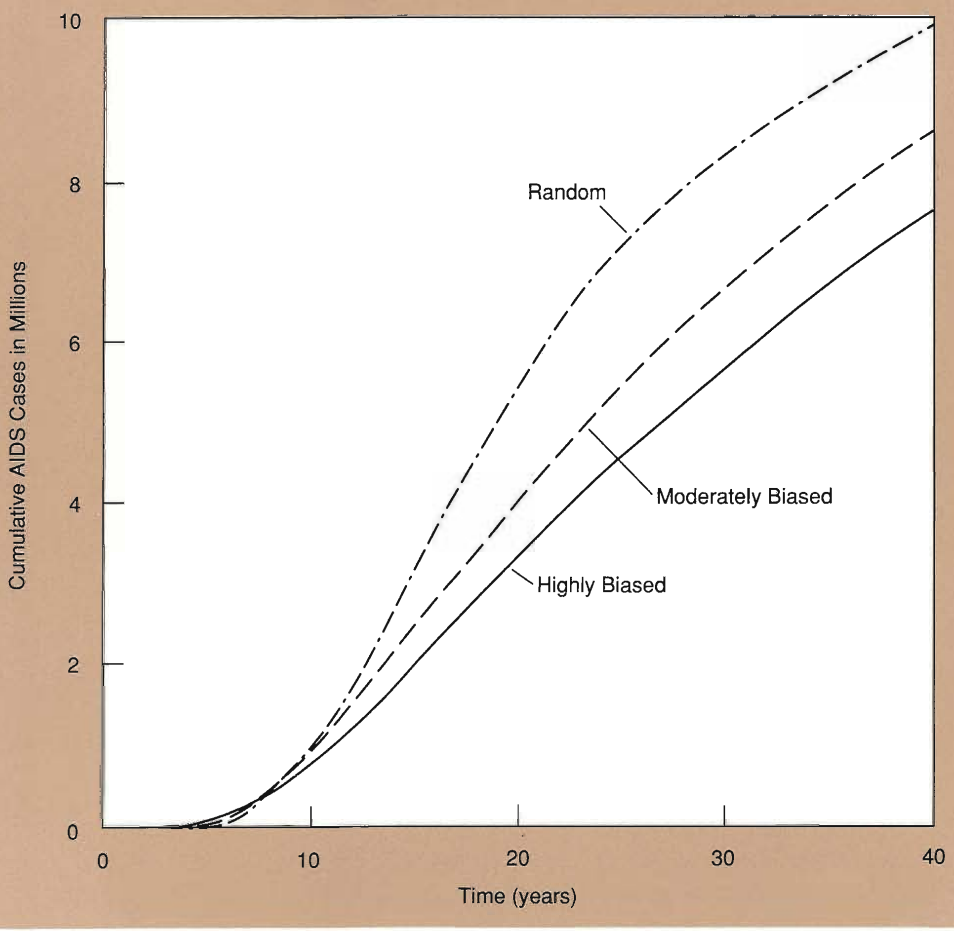
"CUBIC GROWTH" OF BASELINE SOLUTION

Fig. 6. The cube root of the cumulative number of AIDS cases as a function of time for the baseline solution. Although the curve is not perfectly straight, a t^3 growth in the cumulative number of AIDS cases is a good fit to this calculation between $t = 1$ and $t = 9$ years. Thus, despite the many complexities included in the numerical model, its solutions behave quite similarly to the analytic calculation of the main text. Note that the calculated time scales are fixed by the average value we assume for the product $c(r,s)i(\tau)$ and are therefore highly uncertain.



BIASED VERSUS RANDOM MIXING

Fig. 7. Time-dependent behavior of the number infected and the number of AIDS cases for various degrees of mixing among people with different risk behaviors. The baseline calculation (solid line) corresponds to the highest bias, or narrowest range of mixing. As the range of mixing widens, the epidemic changes dramatically. The growth pattern of the number infected appears to change more than that of the AIDS cases partly because of the scale of the plot, and partly because the slow conversion to AIDS smears out the effects of the change in the number infected. More biased mixing produces a more rapid initial growth than does random mixing, but growth slows down as the infection spreads among low-risk people and the total epidemic is smaller than that produced by random mixing. When mixing is random, high- and low-risk people, pass the virus back and forth between them, so an infected person is much more likely to encounter an uninfected person until the whole population saturates.

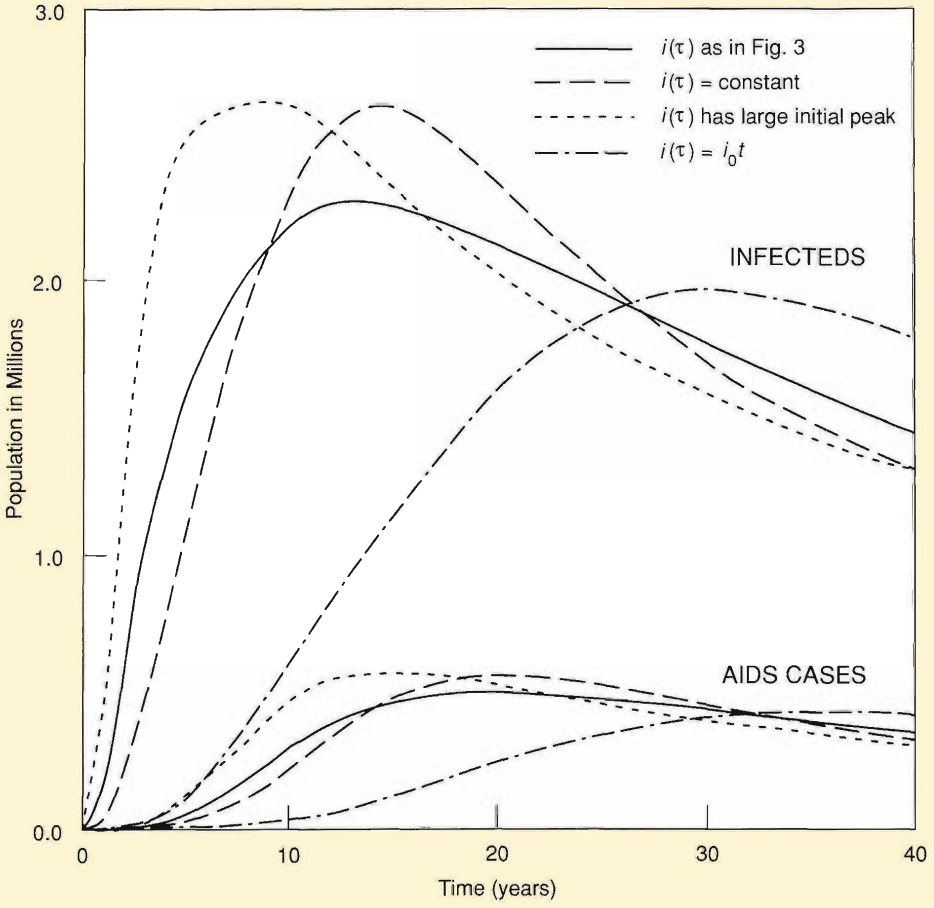


CUMULATIVE GROWTH IN AIDS AS MIXING VARIES

Fig. 8. Cumulative AIDS cases versus time for the calculation in Fig. 7. When mixing is random, the cumulative number of AIDS cases grows exponentially until the entire population reaches saturation of infections. When the mixing is highly biased, the number grows more as a polynomial.

EFFECTS OF VARYING THE INFECTIVITY

Fig. 9. The distribution of number infected $i(\tau)$ as a function of new-partner rate at t = 10 years for the calculations in Fig. 7. This figure demonstrates most dramatically the effects of varying the mixing patterns. When people have a strong bias to mix with others of similar risk, few people of low risk are infected in the early stages of the epidemic. In contrast, when partners are chosen purely on the basis of availability, people of low risk are infected early. The fact that early AIDS cases and early cases of infection were among people with high new-partner rates is evidence for biased mixing in the U.S. population.



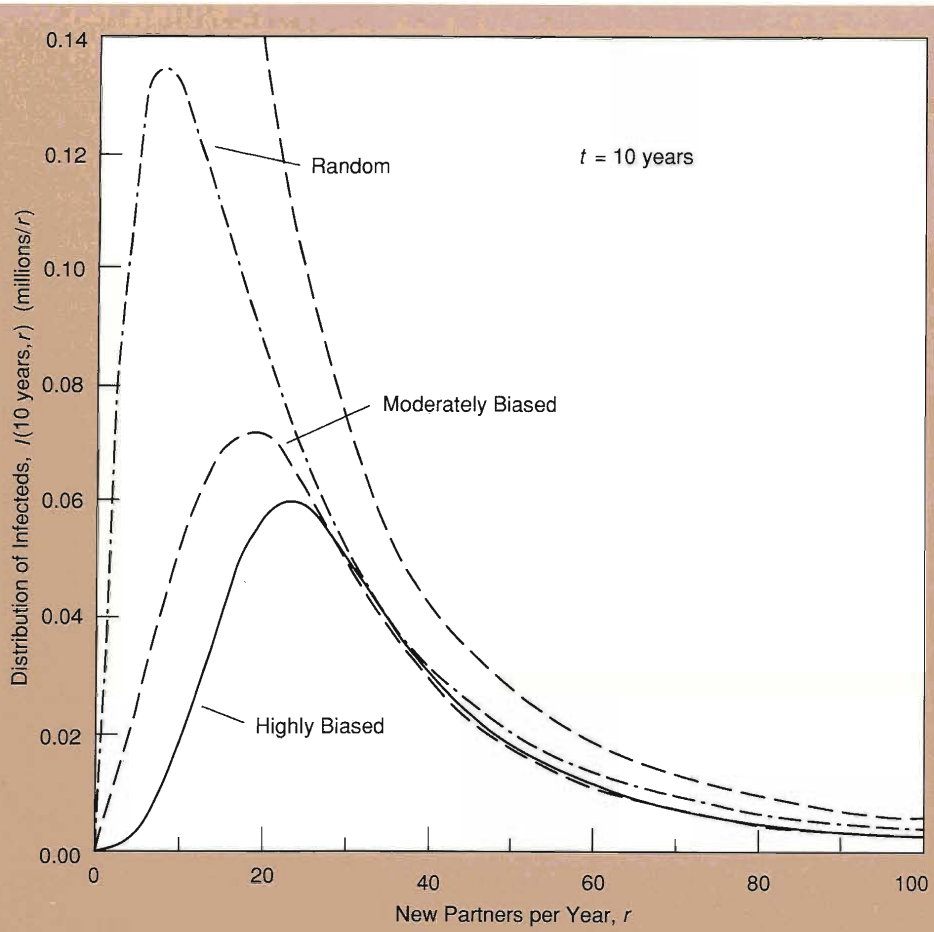
level of mixing among different risk groups. The solid lines show the baseline solution in which the mixing is strongly biased; that is, $f(r,s)$ is an inverse quartic with $\epsilon = 0.01$ (see Fig. 2). The dotted lines show a solution with less bias; that is $f(r,s)$ is again an inverse quartic but $\epsilon = 0.17$ so the curves of $f(r,s)$ versus s for different values of r have much wider peaks than those in Fig. 2. The dashed lines show a solution with no bias; that is, $f(r,s) = 1$ corresponding to random, or homogeneous, mixing. Note that as the mixing becomes less biased, the epidemic starts off slightly later but then grows faster because the doubling time increases at a slower rate.

Figure 8 shows the cumulative number of people with AIDS as a function of time for the three types of mixing. For random mixing, the number of people with AIDS grows nearly exponentially; that is, the doubling time is nearly constant. As the mixing becomes more biased, the number of people with AIDS grows more like a low-order polynomial.

It is worth cautioning that the initial distribution of infecteds, which is arbitrary, can have a significant impact on the early growth of the epidemic, especially if the initial growth rate is low. For the random-mixing case, growth in infections is so low initially that most people getting AIDS in the first 10 years were infected at $t = 0$. Consequently, since those infected at $t = 0$ were distributed linearly with τ, the number of AIDS cases grows as a polynomial during the first 10 years, and only the number infected grows exponentially. After 10 years both the number infected and the number of AIDS cases grow exponentially. For the cases of more-biased mixing, the initial growth in number of infecteds is more rapid, so the initial distribution $I(0,\tau)$ affects the solutions for a shorter period of time. Since our initial conditions are arbitrary, rather than based on knowledge of the earliest stages of the epidemic, the solution transients just described are also arbitrary.

Figure 9 shows the number infected versus risk behavior at $t = 10$ years for each of the three mixing patterns. We see that random mixing not only produces a faster-growing epidemic but also causes the epidemic to reach the low-risk groups almost immediately. Figures 4a and 4b of the main text also illustrate that point. The solution with biased mixing shows a saturation wave of infection traveling from high-to low-risk groups, but the solution with homogeneous mixing shows no such wave. Instead, the majority of those infected are always in the low-risk groups. Since the average partner rates for the earliest AIDS cases and infected homosexuals were high compared to the mean in the general homosexual population, these numerical results support the conclusion in the main text that biased mixing has produced the cubic growth of the AIDS epidemic.

We will now examine the effects of varying the function $i(\tau)$, the infectiousness since time of infection. In the main text we used a constant value of $i(\tau)$, but we also discussed the effects of a variable infectiousness. Here we display four solutions, each of which uses a different function for $i(\tau)$ (see Fig. 10). In all cases the mean infectiousness of an individual over the course of infection is 0.025. The solid lines correspond



INFECTEDS VERSUS RISK AS MIXING VARIES

Fig. 10. Time-dependent behavior of the number infected and the number of AIDS cases for various assumptions about the time-dependence of infectiousness. In these calculations we assign the same value for the average invectivity of any individual over the course of the epidemic and vary only the distribution of infectivity with time. A burst of infectivity just after infection causes the disease to spread very rapidly in the high-risk groups but has less effect as the disease spreads to groups with lower new-partner rates. A slowly rising infectivity several years after the initial burst sustains the epidemic in low-risk groups. With no initial burst of infectivity, but only a slow increase from infection until death, the epidemic initially spreads very slowly, but as more people approach the later stages of infection, the epidemic gains momentum. Without control measures the epidemic may eventually affect as many people as the other examples shown in the figure.

to the baseline solution shown earlier; $i(\tau)$ for that solution is shown in Fig. 3. The dashed lines are the solution when $i(\tau)$ is constant. The dotted lines are the solutions when the infectivity of a person getting AIDS at 8 years has a very large initial peak, then a 4-year period during which $i(\tau) = 0$, and finally a slow increase in $i(\tau)$ up to 8 years since infection. The dash-dot lines are the solutions when $i(\tau)$ has no initial peak, but instead, a person's infectivity increases continuously between the time of infection and the time of AIDS. A large initial peak in $i(\tau)$ produces the fastest-growing epidemic, the absence of an initial peak produces the slowest-growing epidemic, and a constant value for $i(\tau)$ produces an epidemic that is closest to the baseline solution but grows a bit more slowly at first, then somewhat faster, and finally approaches a similar steady state. (Note that the vertical scale in Fig.10 is a blow up of the vertical scale in Fig. 7.) In all cases the growth is "polynomial" in that the doubling times increase continuously. Nevertheless, the shape of $i(\tau)$ has a significant impact on the course of the epidemic.

Without better data for $i(\tau)$, the future course of the present epidemic cannot be estimated. Similarly, adequate data on the mixing patterns among different risk groups is sadly lacking. If nothing else, our risk-based model points out the areas for which more data are needed. We hope that this work will help to guide the data collection and analysis efforts that are now under way. ■

The Seeding Wave

by Stirling A. Colgate and James M. Hyman

Let us assume that our risk-based model is a reasonable description of how AIDS has grown since the time when a member of the highest risk group was infected. In other words, we assume the infection spread as a saturation wave from the highest risk group down through lower and lower risk groups. The question remains—what happened *before* the start of the saturation wave? Did an individual from the highest risk group become infected first and start the saturation wave immediately, or did an individual from a much lower risk (and therefore much larger) group start a slow spread of infection from lower to higher groups prior to the saturation wave? We call a slow spread of infection from lowest to highest risk groups a *seeding wave*. Now, if a seeding wave *were* started, do subsequent seeding events circumvent the slow spread by leapfrogging the infection to the highest risk group, thereby reducing the number infected before the start of the saturation wave?

Here we present a model of a seeding wave consistent with our saturation-wave model of subsequent growth. In particular, the model incorporates the same distribution of risk behavior and assumptions about biased mixing used in our risk-based model. We argue that, provided these assumptions are correct, the seeding wave is a likely scenario for the early spread of HIV infections in the United States. Moreover, the model predicts that the earliest HIV infection occurred in the mid-sixties, a prediction consistent with the first recognized case of AIDS in St. Louis in 1969.

Early Growth. Suppose the first infection in the United States is initiated, say, by either a visitor with HIV or a U.S. person visiting elsewhere. Although these two cases would not be equivalent if high risk of infection is correlated with high rate of travel, we will not consider such correlations here. Rather, we assume that risk of infection can be quantified using a single variable r with its distribution $N(r)$ defined by Eq. 13 in the main article. Since the probability of a person becoming infected is proportional to r, the probability $P(r)$ that at least one individual with risk r or greater becomes infected is given by

$$P(r) \propto \int_r^\infty N(r) r dr = r^{-1}, \qquad (1s)$$

for $r \geq 1$, that is, for the high-risk end of the population defined by Eq. 13.

Hence, the smaller r is, the greater is the probability that at least one individual of risk group r becomes infected despite the lower risk per individual. Also, the most likely case is that the first infected individual was a member of the average group, the group with $r = 1$.

A Simple Numerical Model. We wish to model the progression of the infection to higher risk groups starting with an infected individual close to the average. To help understand this process, we simplify by saying that the kth risk group r_k varies in risk behavior by a factor of 2, that is, r varies from r_k to $2r_k$. Hence, the various groups will have risk behaviors 1, 2, 4, 8... times the average. The number of individuals for $r > 1$ in the kth group (using Eq. 13) is

$$\int_{r_k}^{2r_k} N(r) dr = -\frac{1}{2} N_0 r^{-2} \Big|_{r_k}^{2r_k} = \frac{3}{8} N_0 r_k^{-2}. \qquad (2s)$$

Since the total population (the integral of Eq. 13 from $r = 0$ to $r = \infty$) is $\frac{3}{2}N_0$, our first group with $1 < r < 2$ is one-fourth of the total, and if we restrict ourselves to the homosexual population, one-fourth of the total is one million. Thus, the second group, with $2 < r < 4$, will have $(\frac{1}{4})(\frac{3}{8})N_0$ individuals, or $\frac{1}{16}$th of the total population, or 250,000. The third group, with $4 < r < 8$, will have $(\frac{1}{16})(\frac{3}{8})N_0$, or $\frac{1}{64}$th the total population, and so forth.

We do not believe that people exhibiting a preference for each other are likely to recognize a behavior difference much finer than a factor of 2, and hence, we use this rather crude measure of a group. We also suppose, as a reasonable but unknown example, that the fraction of the time an individual participates in risk outside his group is $F = \frac{1}{4}$. If this fraction is greater or less by a factor of 2, it will change what follows by a factor of 2, but that change is within the accuracy of these estimates.

We next ask how many individuals must be infected in group 1 before a member of group 2 is infected. There are one-fourth as many people in group 2 as in group 1 with twice the risk behavior; that is, the number of each group decreases as $1/r_k^2$ (from Eq. 2s), and they have contacts with other groups only one-fourth of the time. This fraction of out-of-group mixing will be distributed between both higher and lower risk groups.

Let us assume that the fraction is evenly divided between the higher and lower groups and, because of the same bias that leads to group preference, is primarily in the adjacent groups. Crudely then, F can be considered to be a diffusion coefficient. A bias towards only adjacent out-of-group mixing prevents the infection from leapfrogging to much higher groups and circumventing the slow seeding-wave progress.

The seeding wave progresses from group k to the next higher group $k + 1$ when one member of the next higher group is infected. If I_k is growing exponentially, $I_k = e^{(1-F)\alpha r t}$, then the cumulative probability of infecting one member of group $k + 1$, starting at the time when one member of group k is infected, is

$$\begin{aligned} I_{k+1} &= \int_0^t \frac{F}{2}(1-F)(\alpha r_k) I_k dt \\ &= \frac{F}{2} e^{(1-F)\alpha r_k t}\Big|_0^t \\ &= \frac{F}{2}(I_k - 1), \end{aligned} \tag{3s}$$

where the factor $\frac{F}{2}$ is needed because only one half of the out-of-group mixing pertains to the higher risk groups. The remaining half augments the growth rate of the next lower risk group.

Since $I_{k+1} = 1$ when the seeding wave progresses by one group, I_k at this transition becomes equal to $\frac{2}{F+1}$. Therefore, in our example (for which $F = \frac{1}{4}$), nine members of a group must become infected before a member of the next higher group becomes infected. The time for this to occur will be

$$t_k = \frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha r_k}. \tag{4s}$$

Thus, the speed of the seeding wave is $dk/dt = 1/t_k$. The remaining time for the seeding wave to go from group k to the highest risk group at $k = m$ is

$$t_{km} = \sum_k^m t_k, \tag{5s}$$

or

$$\begin{aligned} t_{km} &= \frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha} \sum_k^m \frac{1}{r_k} \\ &\simeq 2\frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha}, \end{aligned} \tag{6s}$$

for $m \gg k$. That is, the sum $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \cdots \simeq 2$ after even just a few terms. Thus, for most of the groups with $k < m$, the remaining time needed for the seeding wave to move through essentially all groups (that is, all but the few of highest risk) is just double the time to infect the adjacent next higher group. Now, each of these seeded groups is growing exponentially so that, as the time increases from, say, t_k to $2t_k$, the number infected in the kth group increases from I_k to I_k^2. Thus, the number of individuals infected in each group at the time the seeding wave ends, $t = t_m$, will be the square of the number infected when the next higher adjacent group is seeded with one individual, or $(\frac{2}{F} + 1)^2$. That is,

$$\begin{aligned} I_k(t_m) &= \int_0^{t_m} e^{(1-F)\alpha t} dt \\ &\simeq (\frac{2}{F} + 1)^2 \quad \text{for } k \ll m. \end{aligned} \tag{7s}$$

Since the seeding wave progresses through m groups and each group has one-fourth the members of the next lower risk group, $m = \ln \frac{N_0}{2} / \ln 4 \simeq 11$, for a total host size of 4 million, or $\frac{3}{2}N_0$. Thus, the maximum number likely infected at the start of the seeding wave is $m(\frac{2}{F} + 1)^2 \simeq 860$. Of course, the out-of-group mixing fraction F is only poorly estimated, and a factor of two larger or smaller value for F implies a range of 270 to 3000 infected before the start of the saturation wave. Although these estimates cover a wide variation, they are upper bounds on the number infected before the start of the saturation wave. As mentioned above, leapfrogging would circumvent the seeding wave and reduce the number infected prior to the start of the saturation wave. Moreover, these upper bounds are not inconsistent with the prediction in the main text that the size of the infected cohort before the start of the saturation wave, namely I_0 in Eq. 24, is small.

This very simple description of the initial spread of infection opens up a number of questions. (1) What is the likely time when the first individual

was infected and, hence, later became the likely first case of AIDS? (2) Is the predicted risk behavior of the early cases of AIDS, inclusive of the seeding wave, consistent with the high mean risk behavior of the early AIDS cases observed by the Centers for Disease Control (CDC)? (3) What is the probability that the whole process of group-to-group progression is circumvented by one high-risk individual becoming infected early in the seeding process? (4) Is the seeding process consistent with our perception that all major demographic groups participated in a near simultaneous start, that is, synchronization of the saturation wave?

Infection Time. We would like to associate a real time with the time step t_k of Eq. 4s and then take the sum $\sum_1^m t_k$ as the total, or maximum likely, time of the seeding wave. This then becomes the *maximum* time prior to 1979.2 that the first person in the United States was likely to have been infected.

In the seeding-wave process, the growth rate of any given group is $(1 - F)\alpha r_k$, where the factor $(1 - F)$ recognizes that out-of-group mixing is not balanced by equal and opposite in-group mixing. We now use the current growth rate of the group at the front of the saturation wave to calibrate $(1-F)\alpha$. In this way, we derive a very rough estimate for the maximum time of the seeding wave.

Figures 2 and 3 of the main article indicate that, at the time 1988.2, the homosexual fraction was approximately 65 per cent of our estimated one million infected, which is 650,000 infected, or $\frac{1}{6}$th of our estimate of the total number of active homosexual population of 4 million. This estimate places the presently infected population partly in group 1 with all higher groups near saturation. The total population already infected in the higher risk groups is $\frac{4}{3}N_2$, or roughly 330,000 (Eq. 2s). Thus, about the same number must be infected in group 1 so that the total is 650,000.

It has required 9 years for the seeded fraction of 81 individuals in group 1 to grow to 320,000, which gives a growth rate of $(1 - F)\alpha = \frac{1}{9} \ln \frac{320000}{81} = 0.92$ per year or a doubling time of 0.75 years. Thus, the apparent growth rate for the total epidemic, which must be averaged over both group 1 and all higher risk groups—groups that, by now, are almost saturated, gives a doubling time that is roughly twice as large, or 1.5 years. This doubling time is to be compared to the present doubling time for infection predicted by our saturation-wave model, which, at $t + 2 = 9$ years, is $(\frac{1}{I}dI/dt)^{-1} \ln 2 = 0.69t/2 = 3$ years. The three-year doubling time corresponds to a two-year doubling time for AIDS, in agreement with present CDC estimates of 1.75 years. Thus, our saturation-wave model is consistent with the CDC data but inconsistent with the simple seeding-wave growth by a factor of two. One source of discrepancy is our incomplete treatment of the effects of out-of-group mixing. We therefore estimate that the growth rate in group 1 is bounded by a doubling time of 0.75 to 1.5 years.

In Eq. 2s we have neglected group 0 $(0 < r < 1)$ with 3.3 million individuals. The first individual infected is equally likely to be in group 0 or 1 because the average value of $N(r)r$ is approximately the same for both groups. We neglected group 0 to simplify the seeding-wave calculation, but since our estimates for the doubling time are too short, we must now recognize that the initial infected individual most likely had a lower mean risk than group 1 and that the mean growth rate is between the growth rate of two groups. As a rough approximation, let us say that the mean growth rate is lower by a factor of $1/\sqrt{2}$. Then the doubling time of the combined group average will be $0.75\sqrt{2}$ to $1.5\sqrt{2}$ years, or 1.1 to 2.2 years. This then becomes a rough estimate of the doubling time of the seeding wave.

First Infection. The date for the beginning of the saturation wave or power-law (t^2) growth of infection was 1979.2 (Eq. 24). But the seeding-wave model suggests that the first infection in the United States may have occurred $\ln((\frac{2}{F} + 1)^2)/\ln 2 \simeq 6$ doubling times earlier, or 7 to 14 years earlier. The date of the first infection thus may fall somewhere between 1972 to 1965, earlier than has previously been estimated. Thus, the singular case of a teenage boy in St. Louis who has now been identified as having died of AIDS in 1969 is consistent with our seeding-wave picture if he was infected up to five years before developing AIDS. The existence of this case of AIDS in 1969 implies a slow growth of the number infected before the start of the saturation wave.

Mean Risk Behavior. We wish to confirm that our model of the seeding wave, which starts in relatively low-risk groups, is consistent with the CDC observation that most early cases of AIDS were among high-risk individuals. The mean risk behavior of those developing AIDS can be calculated using a convolution integral similar in structure to Eq. 26, but one emphasizing risk rather than time since infection. However, here we are really interested in risk behavior versus time and the absolute number of cases of AIDS, because it was the occurrence in 1981 of roughly 50 AIDS cases in a relatively short period of time (approximately 6 months) that caused the recognition of an epidemic.

We next must define high- and low-risk behavior in terms of our seeding-wave model. The new-partner rate of the homosexual population in London SDT clinics (Fig. 5 in the main text) has a mean of roughly 24 partners per year.

We associate this new-partner rate with group 1 of the seeding-wave model. Group 2 would then have a mean rate of 48 new partners per year—well within the CDC definition of extremely high risk behavior. Thus, moderate or low risk behavior is restricted to groups 1 and 0 with doubling times of 1.1 to 2.2 years and 0.8 to 1.6 years, respectively. By 1979 these two groups would each have infected $(\frac{2}{F}+1)^2 \simeq 81$ individuals. Two years later, in 1981, the combined groups would be producing AIDS cases at a rate of 6 per cent per year, that is, $0.06 \times 2 \times 81$, or 10 cases per year. The total cases for 1981 was several hundred, so these 10 additional moderate-risk cases would, by comparison, be negligible. Thus, we believe that the seeding-wave model is consistent with the CDC observation that high-risk behavior was strongly correlated with the AIDS cases at the start of the epidemic.

Bypassing the Seeding Wave. Of course, this slow growth for 7 to 14 years in group 1 could have been bypassed by one member of any group with $r \geq 2$ becoming infected at the beginning. The probability of this happening per infection in group $k-1$ is, for each group, proportional to $N_k r_k \propto F/(2r_k)$ per group, discounting mixing biases. Therefore, the probability that at least one member of higher risk becomes infected, exclusive of the seeding wave, becomes

$$P = \frac{F}{2}\sum_{2}^{m}\frac{1}{r_k} \cong \frac{F}{2}. \qquad (8s)$$

That is, when one member of group 2 becomes infected, it is equally likely that a member of any higher risk group will become infected, and then the remaining time to the start of the saturation wave becomes negligibly small. This effect would reduce the time for the start of the saturation wave by a factor of 1/2 or less, or just the time to infect one member of group 2, which is within the error of our estimates. On the other hand, if we wish to preserve this factor of 2, we must require that F is a function of r_k/r_{k+n}. A bias function as weak as $F \rightarrow F/\ln(r_{k+n}/r_k)$ guarantees that the roughly 100 out-of-group infections likely to occur during the course of the seeding wave have a small probability of being in the highest risk group ($k+1 \leq m$). Otherwise, infection will leapfrog to reach saturation in one-half the time.

The seeding time would likewise be shorter if the infectious source (in another country, for instance) grew rapidly enough to cause many infections in group 1 and, hence, at least one infection in higher risk groups. There is also the possibility that the infection started and died out several times in groups 0 and 1 before starting the seeding wave. This possibility is equivalent to saying that the net reproductive rate of the disease is very close to unity in these low-risk groups, that is, that a given infected individual infects only one other in the mean time of 10 years to AIDS and death. Because of the arguments in the main text concerning the probability of infection per sexual contact and the equivalence of new-partner rate and contact frequency, we believe the net reproductive rate in the homosexual population was large, and thus the seeding wave started with the first infection.

Synchronization of Risk Populations. We ask if either the slow seeding wave or the singular high-risk initial case of infection makes any difference to the saturation-wave model. For sexual preference and race (Figs. 3 and 4 in the main text) as well as regional and age populations (not shown), the cube root of the cumulative number of cases is nearly linear for $t \geq 1982.5$. These curves extrapolate to zero at approximately the same time with a maximum delay of half a year. This result means that all these subpopulations had to be seeded with at least one high-risk member infected within this time interval. The number infected after half a year (using Eq. 24) becomes roughly 2000 or 3000, all within high-risk categories and with or without the seeding wave of initial infection. We then ask what the probability is that any population selected does not have one member within the 2000 to 3000 initially infected high-risk groups. This depends upon the social isolation, but for a subdivision that creates 10 or more categories, no one population is likely to have less than 100 to 150 members in its high-risk group. Thus, isolation would have had to been very strong, such that none were infected. The observed synchronization of the subpopulations seems reasonable and is independent of whether a seeding wave or single high-risk infection started the epidemic.

In summary, we have described a plausible process by which the initial infection of HIV spread in the various risk populations of the United States. Initially, an average individual was infected from sources unknown, but the infection then grew in a peer group until the number infected and the probability of out-of-group mixing caused the infection to jump to a higher risk group. In this fashion, a seeding wave of infection steadily climbed to the highest-risk individuals. The rapid growth among these highest-risk individuals caused all of them to be rapidly infected, resulting in the start of saturation-wave growth for the whole population. The total number infected in the initial seeding wave is strongly dependent upon the out-of-group mixing fraction, but reasonable estimates indicate that the number infected by the seeding wave would be small enough, less than several thousand, to leave the later saturation-wave growth intact. The earliest known case of AIDS in the U.S. in 1969 is consistent with this picture. ■